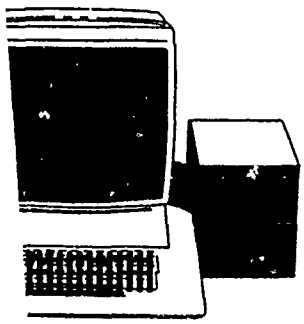
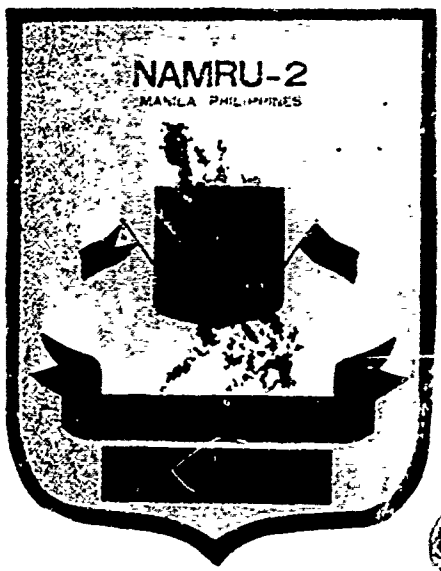
(13)



A GUIDE TO HANDLING BIOMEDICAL DATA*

by

Richard See



A SPECIAL PUBLICATION

OF THE

U.S. NAVAL MEDICAL RESEARCH UNIT NO. 2

MANILA, PHILIPPINES

NAMRU-2 - SP-46

1982

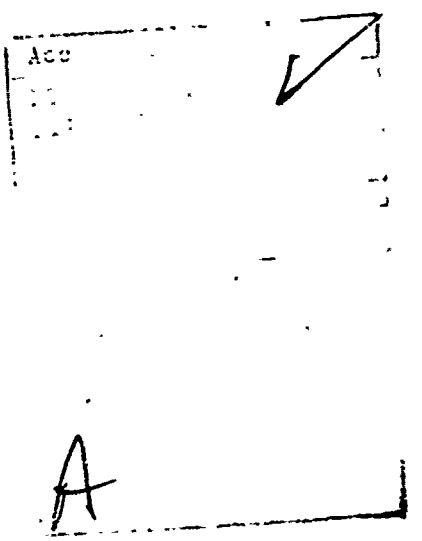
A GUIDE TO HANDLING BIOMEDICAL SURVEY DATA*

by Richard See, CDP

U.S. Naval Medical Research Unit No. 2,

Manila, Philippines

* Presented at the National Workshop on Biomedical Research Methodology, Beijing, People's Republic of China, May 4-15, 1981.



I. Importance of Proper Data Handling Procedures in Medical Research

Almost all medical research today involves the gathering, processing, analysis, and presentation of data. In some studies the data may be incidental and limited in quantity. In others, they may be voluminous and critical to the entire project. In particular, epidemiological studies "are concerned with explaining the occurrence of disease in human populations and with exploiting explanations discovered for the development of methods to protect man against disease." (1) Medical surveys are a key step in most epidemiological studies and usually involve large amounts of data. For this reason the discussion which follows will focus on the role of data handling in medical surveys in order to illustrate typical procedures which may be used to help ensure the validity of the data and the conclusions drawn from them.

Successfully completing a research project is like walking a mountain path. There are many points at which one may fall off the path, or lose one's way but perhaps only a few ways to reach the goal. Almost all medical investigators are aware of the need for maintaining the proper level of scientific methodology with regard to the purely medical aspects of their project. For, if not, their project would probably not have been approved to begin with. Many researchers, however, become less careful when it comes to handling the data. They may hand over most of the data to an assistant or clerk, or else to a computer center, with only limited instructions as to what to do. The same scientist who checks and double-checks the medical aspects of the project may assume that the data will take care of themselves. However, people responsible for computer work have come to know the danger of such an assumption.

In the computer field, and in engineering in general, where things must work properly, or not at all, the tendency for things to go wrong is widely recognized. This tendency is often referred to humorously as "Murphy's Law" and is widely acknowledged by computer professionals. What is involved here, of course, is the level of complexity of computers and data processing has reached the point where the probability of at least one aspect failing is very large. This "law" will undoubtedly come to be better known among medical scientists as they become accustomed to the difficulties that can and will arise in handling large amounts of data.

What are the possible consequences of improper handling of data? First, improper data handling may invalidate an entire project. If this is discovered early enough, it may be possible to go back and remedy the damage by correcting the data, or by collecting a new set of data.

If errors or problems with the data are discovered too late, the entire project may be aborted, due to the impossibility, in many cases, of going back to undo the damage. Worst of all, for the medical profession, improper handling of the data may render the results of the project invalid, without this coming to the attention of the investigator, who may assume that all is well. In some cases, invalid results may be so difficult to identify that they may be published and mislead the medical community. It is to avoid such problems and misfortunes that everyone involved in medical research should become familiar with proper methods of data handling.

II. Role of Computer in Handling Survey Data

A. Functions

1. Record Keeping. One of the ways a computer can be of great help in research is in record keeping. Handwritten records may be hard to read, hard to duplicate, and easily misplaced. If research data can be maintained in a computer system or on punched cards, printed output can be produced which can be stored in several places and in many forms for safety and for dissemination. In many cases, this ability of the computer to serve as a "printing press" for the dissemination of data can be its most valuable function. It should be noted that local officials cooperating in medical surveys often appreciate receiving clear computer printouts of the results, especially when promptly received, and that this is one way in which the survey team can partially repay the personnel who cooperated on the local level. This courtesy helps to assure future cooperation in case further surveys are needed.

2. Error Checking The use of a computer can facilitate error checking and correction, thru listing and tabulation of the raw data. It can also eliminate recopying of data, which so easily leads to new errors. With a computer, such transfers of data from one form to another can be made mechanically, with no further addition of error.

3. Statistical Analysis. Once the data have been entered into a computer system or onto punched cards it becomes possible to carry out repeated statistical analyses without further manual handling of the data. Many programs are available for this purpose in a variety of computer languages.

4. Tables and Graphs. After the research has been completed and is ready for publication or reporting, the computer can again be of assistance in preparing tables and graphs. Many modern computers can produce tables and graphs of sufficiently high quality as to be suitable for publication without further work.

B. Patterns of Computer Utilization

1. Direct Use of Computer by Researcher. If a researcher (a) has a computer available, (b) knows how to use it, (c) has statistical programs available, or can write them himself, (d) can enter the data himself, and (e) understands enough statistics to be able to properly interpret the results, then he should be able to do his data analysis with the aid of the computer without help from anyone else. However, it is rare that all of these conditions are satisfied at once, in a single person. Assuming that a computer is available, the researcher may need assistance in any one of the four other areas from (b) to (e). He may need someone to operate the computer for him, or to prepare programs for him. If he cannot type, he will need someone to keypunch the data on punch cards or to enter the data directly, using a terminal. If he doesn't understand the statistical results, he will need to consult a statistician before he can interpret them.

2. Use of a Computer Center. If the researcher has a computer center available which can assist him, he may be able to get the needed data processing work done for him, up to the point of obtaining some statistical output from the computer. If he understands statistics and can tell the computer center personnel exactly what he needs, and can interpret the final results, he can then make good use of the computer. However, unless the computer center is familiar with statistical work and has the necessary programs, he will probably find it very difficult to get the necessary work done in a satisfactory manner.

3. Cooperation with Statistician. Unless the researcher has a background in medical statistics, cooperation with a statistician is essential to the success of any project involving extensive data analyses. The statistician should be able to help with the entire project, from the design through to the final conclusion. The statistician will also be able to ensure the integrity of the data by checking for completeness and accuracy at each stage. For successful data processing and analysis, the statistician must be able to work with the computer center in order to obtain the desired results. Often, special computer programs are needed. Even if the statistician cannot himself do computer programming, thanks to his mathematical training he will be able to collaborate with a computer programmer to make sure that the correct statistical manipulations are performed and that the results are meaningful.

C. Criteria for Deciding to Use Computer

1. Personnel. Persons having all of the capabilities mentioned above must be available before a decision can be made to use a computer. The following skills must be available: Data Entry Operator (keypunch or terminal), Computer Programmer (if needed programs are not available), Computer Operator, and Statistician. In addition, someone must be given the responsibility for overseeing the proper flow of the data from step to step, and the statistician should ideally play this role. If he does not, then either the researcher or someone he appoints must take this responsibility, or else the whole project may fail.

2. Volume of Data. As a rule, if more than 100 cases are to be included in a study and several variables are involved, the use of a computer should be considered. Even with fewer cases, if there are very many variables, or if a great deal of repetitive analysis is required, the use of a computer may be advantageous. However, one should not use a computer without adequate justification, because of the considerable cost of preparing the data for use by the computer, which may exceed the advantage gained, in some cases.

3. Planning for Computer Use. Computers and computer personnel are a resource, and, like any resource, are limited. Therefore, it is important to plan from the very beginning of a project for the use of data processing, if it is justified. Normally, before a research project can be carried out, a plan, or a proposal, must be prepared covering all anticipated costs and resources needed. In particular, data processing costs and resources which will be needed must also be included in any research plan or proposal. To show that such plans are realistic it is necessary (a) to show why a computer

is needed, (b) what resources will be required, and (c) to indicate their availability and cost, including personnel, equipment, programs, and supplies.

III. Stages of a Medical Survey Project

A. Goal Formulation and Planning

Before any scientific project is carried out, its goals must be thought out and clearly stated. One of the reasons why a clear statement of project goals is important is that the organization providing the support for the project must evaluate the probable value of achieving the stated goals and compare this with the estimated total cost. Although many of the judgments in such an evaluation are subjective, they nevertheless permit a comparison of competing projects and may lead one project to be supported rather than another.

B. Planning and Proposal Preparation

A medical survey is nonexperimental and observational in nature. Therefore, the timing and the selection of the population to be surveyed are of great importance, because there is little or no room for flexibility or revision of procedures once the data gathering has begun. Usually, the purpose of a survey will be to learn more about a particular disease or diseases and the design of the study will depend heavily on what is already known about the disease, its identification, and any variables believed to be related to its occurrence. At this stage it should already be possible to estimate the number of cases to be observed and the number of variables to be recorded. Therefore, it should also be possible at this time to decide whether or not the use of computer will be required for the volume of data involved. If it is determined, for example, by means of statistical estimates of sample size and the number of variables required, that a computer should be used to assist in the processing of the data, then this justification should be included in any plan or proposal which is developed for the project. Even if the amount of data required is small, the complexity of the data processing required may be sufficient to justify the use of a computer.

C. Study Design

From the point of view of the data processing work, the important steps in the study design are the following:

1. Definition of Case. (What constitutes an individual?)
2. Identification of Case. (Procedure for assigning ID numbers, or identification numbers.)
3. Identification of Variables to be Recorded in the Field.
4. Identification of Variables to be Recorded after Survey (Laboratory Results).
5. Specification of Statistical Procedures to be Used in Analyzing Data.
6. Forms Design.
7. Practical procedures for data recording, data entry on forms, data entry (keypunching), data checking (verification of correctness ofdata entry), checking for internal consistency (examination of outliers and statistical abnormalities).

Economy in data gathering is just as important as completeness. No important variables should be overlooked, but at the same time, each variable to be recorded should be scrutinized to make sure that it is really necessary. Previous studies in similar locations or a pilot study in the same area, if necessary, can help in deciding just what variables should be recorded.

D. Carrying Out the Survey

If the planning for the data gathering has been done well, the data collection during the survey will proceed smoothly and accurately. The assignment of case identification numbers (IDNO) for each case will have been clearly arranged and each record and sample (blood, stool, etc.) will be identified with the same IDNO. Advance arrangement with local officials will have been made to obtain accurate population data, including occupation, ethnic origin, and family membership, if relevant. If the entire population is not going to be surveyed, then the statistical sampling procedures must be worked out in advance.

If needed, ecological information on various aspects will also be obtained at the time of the survey, or in advance, if possible. It is often impossible or impractical to obtain some of this information after leaving the survey area. Exact geographical coordinates of the locations surveyed should be established with the aid of maps brought with the survey team. Topographical maps, used at the survey site, are preferable to aid in the establishment of the exact location and the correct altitude above sea level in meters. Accurate place names and names of administrative districts, such a country, and smaller, should be determined and recorded while at the survey site. Weather data, if appropriate for the study should be obtained locally or at nearby administrative centers, and should include seasonal changes. Texts on epidemiology may be consulted to help identify key items of information and to make sure that no necessary information is overlooked (1, 2, 3).

For more ambitious surveys, many types of information may be gathered in order to better understand the transmission of disease in an area. These may include information on vectors, such as mosquitoes, other hosts, such as rats, and patterns of local agriculture, industry, and migration. If such kinds of information are needed, special plans need to be made by specialists in each field.

IV. Codesheets

A. Investigator's Codesheet

Once the overall design of the study has been completed, the investigator will begin to think about the specific elements of data he will need to collect, as outlined above, under Study Design (III. C.). He then should gather these data elements together in the form of a codesheet, which lists all the variables to be recorded and the values they may take on. Assuming that data processing with the aid of punch cards is planned, the investigator may need to think about several types of records, such as those gathered in the field, as contrasted with those created in the laboratory after the survey. Thus, the codesheet may be divided into two or more "record types".

However, if there are not too many variables, it is advantageous and economical to try to put all of the data for each case on a single punch card of 80-columns, i.e. a single record type.

B. Illustrative Example

In order to clarify the nature of the codesheet and the subsequent data processing steps, a Field Study conducted by David T. Dennis will be used as an example (4). In the example given on the following three pages, his hand-prepared codesheet is given. Under "Code No." he has listed the columns on an 80-column form and on the 80-column punch card that will be used for each variable. For example, a 4-digit IDNO is specified in order to leave plenty of room for expansion. This IDNO will be recorded in the first four columns of a form and on the punch card, so 1-4 is specified. For each variable its name and possible values are then indicated, as for Sex in column 9, 1 = male and 2 = male. It is always necessary to anticipate the possible unavailability or loss of information, and the value 9 = unk ("unknown") has been specified to cover this. It is good practice to use the highest possible number of a variable to indicate this unknown value. Thus an unknown age would be coded as "99". Since unknown values are commonplace, statistical programs must be able to accommodate them and a statistical program which cannot accept such unknown values is of little use in survey work.

Some computer programs for statistical analysis permit each variable to be identified with a short name of, say, 5 to 6 letters. For this reason the "Code Name" at the right of each variable has been indicated.

CODE SHEET RECTYPE 1

TDS KUALA KOYAN

Code No.	Information	Code Name
1-4	Registration No. (1001-1250)	REGNO
5, 6	House No. (01-30)	HSENO
7, 8	Age (year) (00-90, 99 = unk)	AGE
9	Sex (1 = male; 2 = female; 9 = unk)	SEX
10	Relationship to Head of Household 1 = HH 2 = wife 3 = child 4 = parent 5 = sibling 6 = other 7 = friend 9 = unk	RELHH
11, 12	Years in village (00-90; 99 = unk)	YRSIN
13, 14	Years out (00-90; 99= unk)	YROUT
15	Healthy? 1 = yes 2 = no 9 = unk	HELTY
16	Bathe River? 1 = yes 2 = no 9 = unk	BATRI
17	Wash clothes in River? 1 = yes 2 = no 9 = unk	WASRI
18	Drink from River? 1 = yes 2 = no 9 = unk	DNKRI
19	Fish in River? 1 = yes 2 = no 9 = unk	FSHRI
20	Sick in belly? 1 = yes 2 = no 9 = unk	SCKBL
21	Bloody diarrhoeae? 1 = yes 2 = no 9 = unk	BLDDR
22	Groin/Axilla pain? 1 = yes 2 = no 9 = unk	KELP

Code No.	Information		Code Name
23	Groin/Axilla abscess?	1 = yes 2 = no 9 = unk	KELAB
24	Swollen extremity?	1 = yes 2 = no 9 = unk	SWNEX
25	Scrotal swelling?	1 = yes 2 = no 9 = unk	SCRSW
26	Scrotal pain?	1 = yes 2 = no 9 = unk	SCRPN
27	Chyluria?	1 = yes 2 = no 9 = unk	CHUIA
28	Malaria?	1 = yes 2 = no 9 = unk	MALIA
29	Elephantiasis?	1 = yes 2 = no 9 = unk	ELEPH
30	Lymphoedema?	1 = yes 2 = no 9 = unk	OEDEM
31	Scarring?	1 = yes 2 = no 9 = unk	SCAR
32	Hydrocoele?	1 = yes 2 = no 9 = unk	HYCLE
33	Thickened epididymis	1 = yes 2 = no 9 = unk	THKEP
34	Elephantiasis Scrotum Breat?	1 = yes 2 = no 9 = unk	SCBRE
35	Hackett Spleen (1-5 ps; 0-5; 9=unk)		SPLN
36	Right liver	1 = yes 2 = no 9 = unk	RTLVR
37	Left liver	1 = yes 2 = no 9 = unk	LTLVR

38	Abdominal venous dilation	1 = yes 2 = no 9 = unk	VENAB
39, 40	Year (80-85; 99=unk)		YEAR
41, 42	Month (01-12; 99 = unk)	MTH	
43, 44	Village Code No. -01 = Kuala Koyan		VILNO
45, 46	Age of infant in months (01-11; 99 = unk)		INFMTH
47	Malaria	0 = neg 1 = pos 9 = unk	MAL
48	Species malaria	0 = neg 1 = P.f. 2 = P. v. 3 = other 4 = mixed P.f./P.v 9 = unk	SPMAL
49	Malaria forms 0 = neg 1 = rings only 2 = gametocytes only 3 = rings and gametocytes 9 = unk		FMMAL
50	Filariasis 0 = neg 1 = pos 9 = unk		FIL
51	Species filaria	0 = neg 1 = B. malayi subperiodic 2 = B. malayi periodic 3 = W. bancrofti 4 = mixed 5 = other	SPFIL
52 - 54	No. mf/20 ul 000 = neg	999 = unk	NOMF
55	Stool for schisto ova	0 =neg 1 = pos 9 = unk	SCHIS
56	Ascaris 0 = neg 1 = pos 9 = unk		ASC
57	Trichuris	0 = neg 1 = pos 9 = unk	TRICH
58	HW 0 = neg 1 = pos 9 = unk		HW
59	COPT 0 = neg 1 = type 1 2 = type 2	3 = type 3 4 = type 4 9 = unk	COPT
80	Rectype No. 1		RETYP

C. Computer Codesheet

Once the investigator has worked out his hand codesheet, it is in a form which can be entered into some computer programs for statistical analysis. In what follows, a set of computer programs, called a computer package, or system, developed by Richard Kronmal, the University of Washington, Seattle, will be used (5). This set of programs is called the "Conversational Computer Statistical System", abbreviated CCSS.

The hand codesheet for the sample survey being illustrated can either be punched on 80-column cards or entered directly into the computer with the CCSS system. Once it has entered into the computer, the computer version of the codesheet can be listed. The following four pages give the computer form of Dennis' codesheet, produced by the CCSS system.

RECORD TYPE 1

PAHANG MALAYSIA FILARIASIS DENNIS

COLUMNS	NUMBER	NAME	VALUES OR RANGE
1- 4	1	REGNO	1001- 1250
5- 6	2	HSENO	1- 30
7- 8	3	AGE	0- 90 ..... 99=UNKWN
9- 9	4	SEX	1= MALE 2=FEMALE ..... 9=UNKWN
10- 10	5	RELHH	1= HH 2= WIFE 3= CHILD 4=PARENT 5=SIBLNG 6= OTHER 7=FRIEND ..... 9=UNKWN
11- 12	6	YRSIN	0- 90 ..... 99=UNKWN
13- 14	7	YROUT	0- 90 ..... 99=UNKWN
15- 15	8	HELTY	1= YES 2= NO ..... 9=UNKWN
16- 16	9	BATRI	1= YES 2= NO ..... 9=UNKWN
17- 17	10	WASRI	1= YES 2= NO ..... 9=UNKWN
18- 18	11	DNKRI	1= YES 2= NO ..... 9=UNKWN
19- 19	12	FSHRI	1= YES 2= NO ..... 9=UNKWN
20- 20	13	SCKBL	1= YES 2= NO ..... 9=UNKWN

21- 21	14	BLDDR	1= YES 2= NO	9=UNKWN
22- 22	15	KELPN	1= YES 2= NO	9=UNKWN
23- 23	16	KELAB	1= YES 2= NO	9=UNKWN
24- 24	17	SWNEX	1= YES 2= NO	9=UNKWN
25- 25	18	SCRSW	1= YES 2= NO	9=UNKWN
26- 26	19	SCRPN	1= YES 2= NO	9=UNKWN
27- 27	20	CHUIA	1= YES 2= NO	9=UNKWN
28- 28	21	MALIA	1= YES 2= NO	9=UNKWN
29- 29	22	ELEPH	1= YES 2= NO	9=UNKWN
30- 30	23	OEDEM	1= YES 2= NO	9=UNKWN
31- 31	24	SCAR	1= YES 2= NO	9=UNKWN
32- 32	25	HYCLE	1= YES 2= NO	9=UNKWN
33- 33	26	THKEP	1= YES 2= NO	9=UNKWN
34- 34	27	SCBRE	1= YES 2= NO	

			..... 9=UNKWN
35- 35	28	SPLN	0- 5
			..... 9=UNKWN
36- 36	29	RTLVR	1= YES 2= NO
			..... 9=UNKWN
37- 37	30	LTLVR	1= YES 2= NO
			..... 9=UNKWN
38- 38	31	VENAB	1= YES 2= NO
			..... 9=UNKWN
39- 40	32	YEAR	80- 85
			..... 99=UNKWN
41- 42	33	MONTH	1- 12
			..... 99=UNKWN
43- 44	34	VILNO	1= KUALA
45- 46	35	INFMTH	0- 11
			..... 99=UNKWN
47- 47	36	MAL	0= NEG 1= POS
			..... 9=UNKWN
48- 48	37	SPMAL	0= NEG 1= P.F. 2= P.V. 3= OTHER 4= MIXED
			..... 9=UNKWN
49- 49	38	FMMAL	0= NEG 1= RINGS 2=GAMETO 3= MIXED
			..... 9=UNKWN
50- 50	39	FIL	0= NEG 1= POS
			..... 9=UNKWN
51- 51	40	SPFIL	0= NEG 1=SUBPER 2=PERIOD 3=BANCRO 4= MIXED 5= OTHER

			..... 9=UNKWN
52- 54	41	NOMF	0- 998
			..... 999=UNKWN
55- 55	42	SCHIS	0= NEG
			1= POS
			..... 9=UNKWN
56- 56	43	ASC	0= NEG
			1= POS
			..... 9=UNKWN
57- 57	44	TRICH	0= NEG
			1= POS
			..... 9=UNKWN
58- 58	45	HW	0= NEG
			1= POS
			..... 9=UNKWN
59- 59	46	COPT	0= NEG
			1= TYPE1
			2= TYPE2
			3= TYPE3
			4= TYPE4
			..... 9=UNKWN
80- 80	47	RT	0= RT1

V. Data Recording

A. Preparatory Work and Form Design

The codesheet should be carefully checked by everyone involved to see that the necessary variables and their proper ranges have been included. When in doubt it is best to allow one more digit that can be used in case of need for a variable. In general, for surveys, one might as well allow 4 digits for the identification number, in case the number exceeds 999.

Age of young children is often a problem and the solution in the codesheet example in the previous section illustrates a good practical solution, namely, record age in full years for everyone. Then, record age in months from 0 to 11, or higher, for infants. Note that months should be coded as unknown = 99 for all adults, to avoid confusion when looking at the variable INFMTH (Age of infant in months). This was not done in the sample, but was taken care of by noting that no child in this particular survey was less than one month old. In the present case, the unknown value for INFMTH could better have been taken as 0. This kind of change in the codesheet after preliminary processing is quite common.

Another comment on unknown values that should be made is that "unknown" to the statistician merely means unavailable. For example, an age could have been known, but then blurred to become unreadable. The CCSS system used in the example, as well as many others, permit only 1 unknown value for a variable. However, some investigators feel the need to indicate in detail the reason why each missing or unavailable value is not present. For example, when blood samples are used up before all testing is completed, the comment is then QNS = Quantity Not Sufficient. If the investigator really wishes to record this information, he should be encouraged to use an additional variable rather than try to use 2 or more "unknowns" for the same variable.

Once the codesheet has been well worked out, a form should be designed which will permit easy and accurate recording of the data in the field. It is often helpful to have the identification numbers printed on the forms in advance. A decision must be made as to the type of form to be used. For some basic population data, it may be desirable to record 30 to 40 cases per sheet. However, for other types of data, it may be best to record all the data on a single case on one sheet.

B. Log Books

On field studies log books are often used. In this case the format of the data recorded should be drawn in columnar form in the log book before the study is carried out. Log books, if bound, have the advantage that they form a single unit, so that individual sheets will not be lost. On the other hand, they are harder to work with after the field study is completed. Photocopying machines may be used to copy data from log books for further processing, but if the writing is not clear, or if the format is not well laid out, the copy may be illegible. The choice will depend on the situation and on the nature of the data to be gathered.

VI. Data Transcription

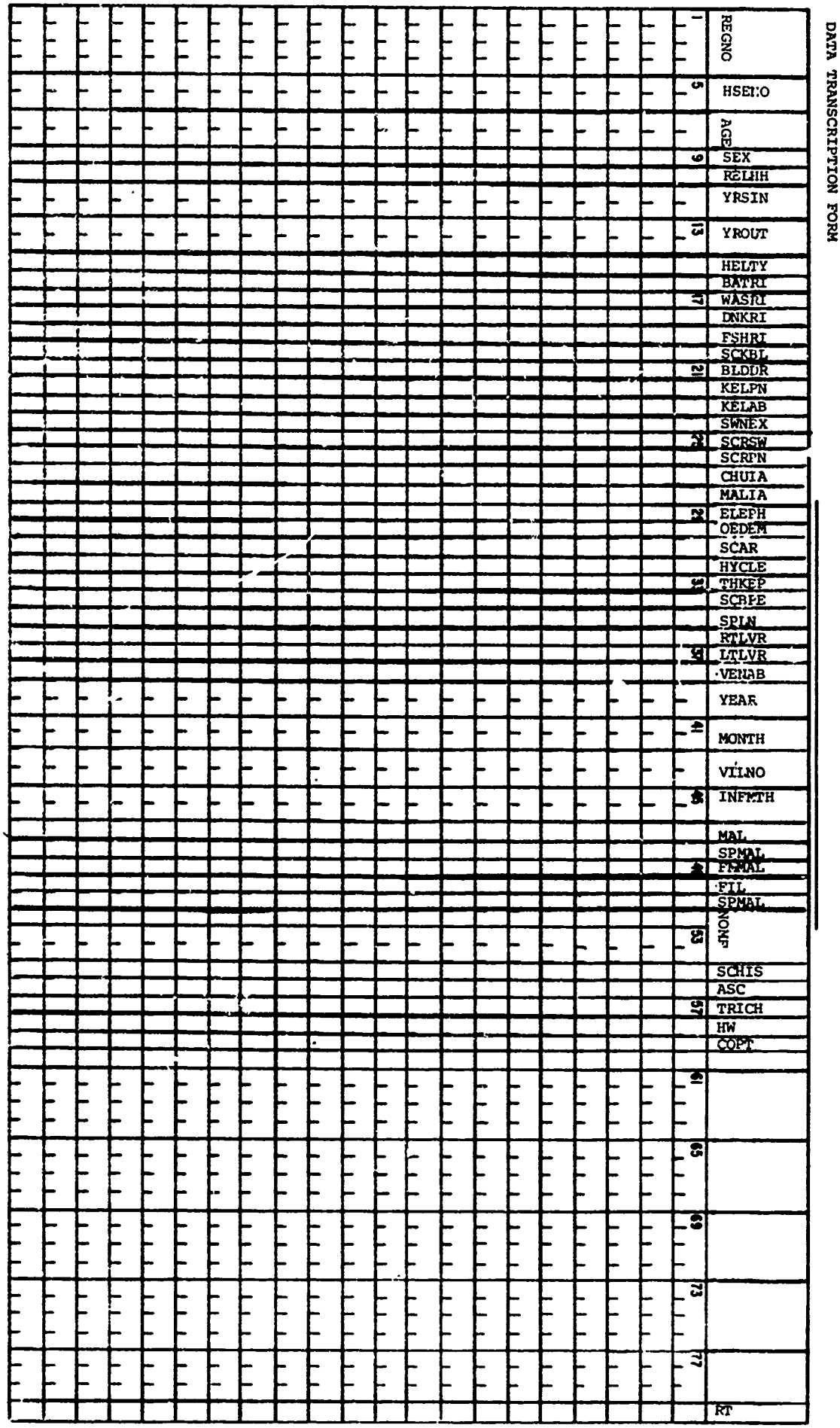
A. Form Design

Assuming that a computer is to be used to process the data, it will usually be necessary to transcribe the data from the field log or from the field data sheets onto special sheets used for data entry. Since the data entry operator will normally have no understanding of the meaning of the data, he or she will be simply transferring numbers, or even digits from the sheets presented into a data processing medium. 80-column punch cards are usually used to record the data for computer use. Therefore, a form with 80 columns must be used to transcribe the data for the punching or entry by the operator. Each line of the 80-column form will represent a case, and there may be 20 or more lines or cases per sheet. If possible, vertical lines should be drawn on the 80-column sheet indicating where each variable starts and ends. The variable name should be indicated at the top of the sheet, over the columns assigned to the variable in the codesheet. The illustration on page 19 shows the design of a data transcription form for the illustrative example.

B. Transcription

Someone familiar with the data and its significance should transcribe it from the field data sheets or from the field logs onto the data transcription form. The transcription should be checked as it is done and made as free from error as possible. In particular, blanks should not be left where the unknown codes for variables specify something else. Many computer statistical systems are based on the FORTRAN language, which has the peculiarity that a blank is interpreted as 0 (Zero), which may have a different meaning entirely.1

DATA TRANSCRIPTION FORM



VII. Data Entry

If punch cards are to be used to enter the data, the keypunch operator will operate the keyboard of the keypunch to prepare 80-column punch cards by typing from the data transcription form (80-column form). It is best that the keypunching be done in a mechanical manner, without any changes or judgments being introduced in the process. However, a highly trained and experienced keypuncher may be able to identify errors and inconsistencies in the data and should then call these to the attention of the investigator by noting the questionable item on the form, even though typing it exactly as transcribed.

Depending on the skill and error rate of the keypunching personnel, it may be desirable to verify the punching immediately, before computer processing. Some modern keypunch machines are dual purpose and also permit a verification process in which all of the material is retyped by another operator, preferably, and the machine compares the typing with the characters already punched in the card. Errors encountered are recorded by a system of notches on the edge of the card. There are also separate machines for verifying card punching accuracy.

For computers using direct data entry through a keyboard, there will be no punched cards and no verification will be possible outside the computer. In this case, the verification must all take place at a later stage.

VIII. Data Presentation and Checking

A. 80-Column List

After the data have been punched on 80-column cards, the cards should be read into the computer and listed in a solid block format, exactly mirroring the digits on the data transcription form. The purpose of this "80-column" listing is to permit the careful checking of all of the data at this stage by comparing the listing with the data as transcribed. In the illustration on the following page the 80-column listing has been made with the additional printing of the variable names, vertically, to save space. A FORTRAN program is available from NAMRU-2 which does this automatically from the cards used to list the computer codesheet. Once this 80-column listing has been thoroughly corrected and relisted, it should be saved as part of the records of the survey, since, if the cards are lost, they can readily be repunched from this listing.

It should be emphasized that the 80-column list is not intended for use in data analysis, since it is extremely hard to read the values of the individual variables from such a list. The 80-column listing for the sample Field Study follows.

JOB 5221 PAHANG MALAYSIA FILARIASIS 18 MAR 81

```
R   H A SRY Y HBWDFSBKKSSSCMEOSHTSSRLVY M V I MSFFSN  SATHC
E   S G EER R EAANSCLEEWCCHALECYHCPTTEE O I N APMIPO  CSRWO
G   E E XLS O LTSKHKDLLNRRULEDACKBLLLNA N L F LMMLFM  HCI P
N   N    HI U TRRRRBDPAESPIIPERLERNVVAR T N M  AA IF  I C T
O   O    HN T YIIIILRNBXWNAAHM EPE RRB  H O T  LL L   S H
                                        H

10010164111648211112222222212229990222800501000000000001119
10020122131606111112222222222229990222800501000000000001111
10030116261006111119222222222222220122800501000000000000119
10040101260100212121292222212222220222800501000000000099999
10050143170142111112292222222222993222800501000000100101119
10060223110221111112221222212212990222800501000001200500110
10070223220221111112212222212222223222800501000000000001100
10080207130205111192222222212222920222800501000000000099990
10090203130201111122222222212222223222800501000111000000110
10100345110540111112222222222222299022800501000000000001110
10110340220535111112222222222222222222280050100001900020l110
10120312130507111119299999999999999999800501000099999999999
10130307130502211129299222212222922222800501000000000001110
10140304130400212122992222122229242228005010011100000l1110
10150301230100212129292222212222120222800501001210000001000
10160442110834111112222222212229990222800501000000000000111
10170428220820211122222222212222220122800501000000000099990
10180409230801211122299222212222220222800501000000000000112
10190407230700211122299222212222222222800501000000000000112
10200404230400212122992222122229232228005010011200000011l0
10210402130200212129299222122229202228005010011200000099990
10220401130100212129299299212222220122800501000000000099999
10230555112728111122222222212222920112800501000000C0099990
10240655112728122222212122212212922222800501000000000099990
10250651222724122221211222212212202228005010000000000099990
10260605160500222221222222122229902228005010000000000099990
10270729112900122222222222212222992228005010000000000001110
10280723220815122222222222212222220222800501000000000001119
10290704130400122222222222212222990222800501000000000000110
10300830110921122222212222212222992222800501000120000501110
10310826220917122222222222212222220222800501000000000001119
10320842260834222222222222212222990222800501000000000001110
10330810230901122222222222212222222222800501000000000001100
10340805130500222221222222122229232228005010000000000001110
10350803230300122222222222222222220222800501000000000001110
10360801230100122222122222222222222222800501001130000000009
10370860140951222222211222212222990222800501000000000001110
10380958111938111112122222212229990121800501000000000001110
10390943221924111112222222222222290222800501000000000001110
10400913231300122222222222222222223222800501000000000001110
10410911131100999999999999999999999980050199999999999999
10420907230700122222222222222222220122800501000000000001119
10430906230600122221222222922222222222800501000000000000100
10440904230400222221222222222222220122800501000000000001110
10450946160442122222222222222222220222800501000001200201100
10460936260432122222211222222212220222800501000001900100110
10470912260408122222212222912222223222800501001110000101110
10480909160405122222222222212222220222800501000000000101110
10490907160403122222222222212222220222800501000000000100110
10500906260402122222222222212222220222800501000000000001110
```

JCB 5221 PAHANG MALAYSIA FILARIASIS 18 MAR 81

```
R   H A SRY Y HBWDFSBKKSSSCMEOSHTSSRLVY M V I MSFFSN  SATHC
E   S G EER R EAANSCLEEWCCHALECYHCPTTEE O I N APMIPO  CSRWO
G   E E XLS O LTSKHKDLLNRRULEDACKBLLLNA N L F LMMLFM  HCI P
N   N    HI U TRRRRBDPAESPIIPERLERNVVAR T N M  AA IF  I C T
O   O    HN T YIIIILRNBXWNAAHM EPE RRB  H O T  LL L   S H
                                         H

105109042604001222222222221222222022280050100000000099990
105209022602001222222222221222222022280050100000000000100
105309011601011222222222222222222022280050100000000001000
105410511103481222222222221221299022280050100000000001110
105510502203471222222222222222222022280050100000000001010
105610162303131222222222222222222022280050100000000001110
105710112303081222222222222222222022280050100000000001110
105810081303052222222222222212222992228005010000000001110
105911421105371222222112222122322980050100000120140111O
106011312205261222222222222212222201128005010000000001110
106111551405502111122222222222222201228005010000000001110
106211101305051222221222222222222212280050100000000001100
106311082305031222221222222222222022280050100000000001110
106411171704132111112222222122222022280050100000120079999O
106511192710091111122222222222222212280050100000120070111 9
106611022302001121222992222122222232280050100999999999999 9
106712401199992222229999999999999999980050100000000001112
106812172202151222222222221222222022280050100000001501110
106912062306001222222222291222222022280050100000000001110
107012042304001222299222222922222022280050100000000001110
107112031303001222299999999999999999980050100000000099990
107212011301001222299999999999999999980050100999999999999 9
107313212207141222222222222222222112280050100000009990111O
107413222604181222222222222222222012280050100000009990111O
107513061306002222222222221222222022280050100000009990111 9
107613032303001222292222222222222022280050100000009999999O
107713011301001222222222221222222022280050100999999999999 9
107813021602001222222222222222222022280050100000000009999O
107913281328001222222222221222299022280050100999999999999 9
108013172317001222222222222222922022280050100000000001110
108113112311002222222222222212222222228005010011300000011OO
108213571120371222222222221222299222280050100000000001110
108314421130121222222222222222292022280050100000000001110
108414382230081222222222222222222022280050100000000001110
108514031303001222222222222222222022280050100999999999110 9
108615291102272222211122221222222022280050100000011025010OO
108715312202291222221222221222222022280050100999999999110 9
108815071302051222221222222222229222280050100000000001110
108915062302041222222222221222222222280050100000000000110
109015001300002222222222221222292022280050109999999999999 9
109115102601091222222222222222222022280050100000000009999O
109216481410381222222111222222222022280050100000000000110
109316482210381222222222221222222012280050100000000000110
109416112311001222222222222222222012280050100000000000110
109516071307009999999999999999999998005010000000009999O
109616031303001222222222222122222211228005010000000000110
109716021302002222222222221222222022280050100000000001109
109816421610321222222222222122229932280050100000000001119
109916352610251222222222222122222012280050100000000001119
110016461610361222222221222121222222228005010099999999999 9
```

JOB 5221 PAHANG MALAYSIA FILARIASIS 18 MAR 81

```
R.   H A SRY Y HBWDFSBKKSSSCMEGSHTSSRLVY M V I MSFFSN  SATHC
E    S G EER R EAANSCLEEWCCHALECYHCPTTEE O I N APMIPO  CSRWO
G    E E XLS O LTSKHKDLLNRRULEDACKBLLLNA N L F LMMLFM  HCI P
N    N    HI U TRRRRBDPAESPIIPERLERNVVAR T N M  AA IF  I C T
O    O    HN T YIIIILRNBXWNAAHM EPE RRB  H O T  LL L   S H
                                          H
```

```
11011622261012122222222222299999999998005010000000000000110
11021641269999122222222222229999999999800501009999999999999
11031741110734122222222222212222993222800501000000000001110
11041830299999999999999999999999999999998005010000019004999 0
11051805299999999999999999999999999999998005010000000000011 6
11061803299999999999999999999999999999998005010000000000099990
11071856113026122222111222212222994222800501000000000001112
11091840193010122222222222222212220222800501000000000000110
11101825260421222222222222222222220122800501001110000000110
11111803160300122222222222212222922222800501000000000001119
12501555243025222222222222222222222222280050100999999999999
```

K. Data Count

Even if the 80-column listing is checked very carefully, it is still possible that some typing or keypunching error could be missed. Therefore, it is useful at this stage to have some kind of analysis, or tabulation, of the characters that have been punched in each of the 80 columns of the punch card.

Since the data format is fixed by the codesheet, every column of the data transcription form and of the punched cards is assigned a specific function. For example, in the sample Field Study data, columns 7 and 8 are used for the age of each case. Therefore, column 7 should contain only the 10-digit, or decade, of the age of eah case, and column 8 should contain all the single digit values. Thus, if there is no one in a survey with age over 69, then the values 7 and 8 should not occur in column 7 of any card. A 9 might occur in column 7 only if some case had unknown age, in which case 99 would be used. In the present example there were no unknown ages, so no 7's, 8's, or 9's should occur in column 7. We would expect in a sample of the present size that ages might well occur ending in all possible digits from 0 to 9, and that is actually the case, as we will see.

A program is in use at NAMRU-2 which prints both the 80-column list and a "data count" table showing all of the characters occuring in each of the 80 columns. The table was named DATAC, and a program of that name was contributed to NAMRU-2 by the University of Washington for use with CCSS. The DATAC program has been modified by NAMRU-2 to print the variable names at the starting column of each variable. Thus, AGE would automatically be printed with column number 7 in the DATAC table. The NAMRU-2 DATAC table is given on the following 2 pages for the sample Field Study.

The DATAC table has two functions: (1) additional checking of the data as punched, and (2) preliminary data analysis. The method of checking has already been illustrated with AGE above. The columns are listed vertically, instead of horizontally, in the DATAC table to overcome problems of space. Thus, the seventh row of the table is labelled "7AGE". (This means that the "age" field begins in column 7.) In the columns from 0 to 9 are tabulated the number of times each character occurred in column 7, say. Thus, by looking at row 7 in the table we can see that 46 cases had age from 0 to 9, 16 cases had age from 10-19, and so on. Only 2 cases were punched as having age in the range 60-69. Thus we have in effect obtained a histogram on age with 10 year intervals. This illustrates the use of DATAC for preliminary analysis. The fact that there are no punches from 7 to 9, or alphabetic punches, or even blanks, is a helpful check for correctness of the punching. Other checking is done in a similar manner.

JOB 5221 PAHANG MALAYSIA FILARIASIS 18 MAR 81

COL NAME	0	1	2	3	4	5	6	7	8	9	BLNK	ALPH	TOTL
1REGNO	0	111	0	0	0	0	0	0	0	0	0	0	111
2	99	11	1	0	0	0	0	0	0	0	0	0	111
3	18	12	10	10	10	11	10	10	10	10	0	0	111
4	12	12	11	11	11	11	11	11	10	11	0	0	111
5HSENO	53	58	0	0	0	0	0	0	0	0	0	0	111
6	5	13	10	16	10	8	14	4	15	16	0	0	111
7AGE	46	16	13	7	17	10	2	0	0	0	0	0	111
8	11	18	15	14	7	10	12	11	8	5	0	0	111
9SEX	0	56	55	0	0	0	0	0	0	0	0	0	111
10RELHH	0	17	14	46	4	0	23	3	0	4	0	0	111
11YRSIN	78	17	6	5	0	0	0	0	0	5	0	0	111
12	15	12	14	12	15	11	5	9	6	12	0	0	111
13YROUT	61	10	15	12	6	2	0	0	0	5	0	0	111
14	43	9	8	4	7	10	5	6	11	8	0	0	111
15HELTY	0	79	27	0	0	0	0	0	0	5	0	0	111
16BATRI	0	29	77	0	0	0	0	0	0	5	0	0	111
17WASRI	0	22	84	0	0	0	0	0	0	5	0	0	111
18DNKRI	0	29	77	0	0	0	0	0	0	5	0	0	111
19FSHRI	0	16	89	0	0	0	0	0	0	6	0	0	111
20SCKBL	0	4	92	0	0	0	0	0	0	15	0	0	111
21BLDDR	0	11	91	0	0	0	0	0	0	9	0	0	111
22KELPN	0	11	80	0	0	0	0	0	0	20	0	0	111
23KELAB	0	7	87	0	0	0	0	0	0	17	0	0	111
24SWNEX	0	3	99	0	0	0	0	0	0	9	0	0	111
25SCRSW	0	0	101	0	0	0	0	0	0	10	0	0	111
26SCRPN	0	0	101	0	0	0	0	0	0	10	0	0	111
27CHUIA	0	0	99	0	0	0	0	0	0	12	0	0	111
28MALIA	0	62	40	0	0	0	0	0	0	9	0	0	111
29ELEPH	0	0	99	0	0	0	0	0	0	12	0	0	111
30OEDEM	0	1	99	0	0	0	0	0	0	11	0	0	111
31SCAR	0	7	92	0	0	0	0	0	0	12	0	0	111
32HYCLE	0	0	96	0	0	0	0	0	0	15	0	0	111
33THKEP	0	1	67	0	0	0	0	0	0	43	0	0	111
34SCBRE	0	0	79	0	0	0	0	0	0	32	0	0	111
35SPLN	66	2	19	11	2	0	0	0	0	11	0	0	111
36RTLVR	0	18	82	0	0	0	0	0	0	11	0	0	111
37LTLVR	0	2	98	0	0	0	0	0	0	11	0	0	111
38VENAB	0	1	98	0	0	0	0	0	0	12	0	0	111
39YEAR	0	0	0	0	0	0	0	0	111	0	0	0	111
40	111	0	0	0	0	0	0	0	0	0	0	0	111
41MONTH	111	0	0	0	0	0	0	0	0	0	0	0	111
42	0	0	0	0	0	111	0	0	0	0	0	0	111
43VILNO	111	0	0	0	0	0	0	0	0	0	0	0	111
44	0	111	0	0	0	0	0	0	0	0	0	0	111
45INFMTH	110	0	0	0	0	0	0	0	0	1	0	0	111
46	109	0	0	0	0	0	0	0	0	2	0	0	111
47MAL	92	8	0	0	0	0	0	0	0	11	0	0	111
48SPMAL	90	9	1	0	0	0	0	0	0	11	0	0	111
49FMMAL	88	6	3	2	0	0	0	0	0	12	0	0	111
50FIL	89	9	0	0	0	0	0	0	0	13	0	0	111
51SPFIL	90	2	5	0	0	0	0	0	0	14	0	0	111
52NOMF	95	0	0	0	0	0	0	0	0	16	0	0	111
53	92	2	1	0	0	0	0	0	0	16	0	0	111
54	80	5	2	0	2	4	0	2	0	16	0	0	111
55SCHIS	83	0	0	0	0	0	0	0	0	28	0	0	111
56ASC	20	63	0	0	0	0	0	0	0	28	0	0	111
57TRICH	5	78	0	0	0	0	0	0	0	28	0	0	111
58HW	14	69	0	0	0	0	0	0	0	28	0	0	111

59COPT	78	2	4	0	0	0	0	0	0	27	0	0	111
60	0	0	0	0	0	0	0	0	0	0	111	0	111
61	0	0	0	0	0	0	0	0	0	0	111	0	111
62	0	0	0	0	0	0	0	0	0	0	111	0	111
63	0	0	0	0	0	0	0	0	0	0	111	0	111
64	0	0	0	0	0	0	0	0	0	0	111	0	111
65	0	0	0	0	0	0	0	0	0	0	111	0	111
66	0	0	0	0	0	0	0	0	0	0	111	0	111
67	0	0	0	0	0	0	0	0	0	0	111	0	111
68	0	0	0	0	0	0	0	0	0	0	111	0	111
69	0	0	0	0	0	0	0	0	0	0	111	0	111
70	0	0	0	0	0	0	0	0	0	0	111	0	111
71	0	0	0	0	0	0	0	0	0	0	111	0	111
72	0	0	0	0	0	0	0	0	0	0	111	0	111
73	0	0	0	0	0	0	0	0	0	0	111	0	111
74	0	0	0	0	0	0	0	0	0	0	111	0	111
75	0	0	0	0	0	0	0	0	0	0	111	0	111
76	0	0	0	0	0	0	0	0	0	0	111	0	111
77	0	0	0	0	0	0	0	0	0	0	111	0	111
78	0	0	0	0	0	0	0	0	0	0	111	0	111
79	0	0	0	0	0	0	0	0	0	0	111	0	111
80	0	0	0	0	0	0	0	0	0	0	111	0	111

IX. Statistical Analysis

A. Listings and Summary Statistics

In order to produce a readable listing of the data for the use of the investigator several things are necessary. First, the variables must be properly spaced. Second, leading zeros must be suppressed, so that age 4 is not printed as 04. Third, each variable should be clearly indicated at the top of the column, using the alphabetic name in the codesheet. Fourth, unknown values according to the codesheet should be replaced by the notation UNKWN, or equivalent for each unknown occurrence of a variable. Fifth, the alphabetic values specified in the codesheet for any variable should be substituted for the numeric values, as FEMALE in place of 2, for example. All of these changes are automatically made by the CCSS system, as can be seen in the following listing of the sample Field Study data. (It should be noted that in the CCSS system, the values of a variable must be either all numeric, or all with alphabetic conversion. It is not possible to specify 0-80 numeric, but 88 = MISSNG and 99 = UNKWN, for example. The UNKWN is always possible, for unknown values, but it is not possible to mix the continuous range 0-80 with the discrete value 88 in the CCSS codesheet.)

Summary statistics for all variables can be produced automatically for all of the data at the time a listing is made. Summary statistics for our sample data follow the CCSS listing. For each variable a number of quantities are given. The sequential number of the variable is given first, followed by its name. Next, the maximum value of each variable (HIGH) and the minimum value (LOW) are given. The number of cases for which each variable is known is then given (KNOWN CASES). This is simply the total number of cases, less the number of cases for which that variable was UNKWN. Finally, the MEAN and S.D. are given. These can be used in statistical analyses. The COEF VAR is simply the S.D. divided by the MEAN then multiplied by 100. It may be helpful in comparing the variation of variables with markedly different means.

One of the most important functions of the listing and summary statistics is for the further checking of the accuracy of the data. Here, errors which arose at a very early stage, such as the copying from a log book, may be found. An unusually large S.D. or COEF.VAR may also be a clue to an error, or outlier. Unexpected HIGH and LOW values may also point to errors and should be checked.

PAHANG MALAYSIA 1GO

REGNO	HSENO	AGE	SEX	RELHH	YRSIN	YROUT	HELTY	BATRI	WASRI	DNKRI
1001	1	64	MALE	HH	16	48	NO	YES	YES	YES
1002	1	22	MALE	CHILD	16	6	YES	YES	YES	YES
1003	1	16	FEMALE	OTHER	10	6	YES	YES	YES	YES
1004	1	1	FEMALE	OTHER	1	0	NO	YES	NO	YES
1005	1	43	MALE	FRIEND	1	42	YES	YES	YES	YES
1006	2	23	MALE	HH	2	21	YES	YES	YES	YES
1007	2	23	FEMALE	WIFE	2	21	YES	YES	YES	YES
1008	2	7	MALE	CHILD	2	5	YES	YES	YES	YES
1009	2	3	MALE	CHILD	2	1	YES	YES	YES	YES
1010	3	45	MALE	HH	5	40	YES	YES	YES	YES
1011	3	40	FEMALE	WIFE	5	35	YES	YES	YES	YES
1012	3	12	MALE	CHILD	5	7	YES	YES	YES	YES
1013	3	7	MALE	CHILD	5	2	NO	YES	YES	YES
1014	3	4	MALE	CHILD	4	0	NO	YES	NO	YES
1015	3	1	FEMALE	CHILD	1	0	NO	YES	NO	YES
1016	4	42	MALE	HH	8	34	YES	YES	YES	YES
1017	4	28	FEMALE	WIFE	8	20	NO	YES	YES	YES
1018	4	9	FEMALE	CHILD	8	1	NO	YES	YES	YES
1019	4	7	FEMALE	CHILD	7	0	NO	YES	YES	YES
1020	4	4	FEMALE	CHILD	4	0	NO	YES	NO	YES
1021	4	2	MALE	CHILD	2	0	NO	YES	NO	YES
1022	4	1	MALE	CHILD	1	0	NO	YES	NO	YES
1023	5	55	MALE	HH	27	28	YES	YES	YES	YES
1024	6	55	MALE	HH	27	28	YES	NO	NO	NO
1025	6	51	FEMALE	WIFE	27	24	YES	NO	NO	NO
1026	6	5	MALE	OTHER	5	0	NO	NO	NO	NO
1027	7	29	MALE	HH	29	0	YES	NO	NO	NO
1028	7	23	FEMALE	WIFE	8	15	YES	NO	NO	NO
1029	7	4	MALE	CHILD	4	0	YES	NO	NO	NO
1030	8	30	MALE	HH	9	21	YES	NO	NO	NO
1031	8	26	FEMALE	WIFE	9	17	YES	NO	NO	NO
1032	8	42	FEMALE	OTHER	8	34	NO	NO	NO	NO
1033	8	10	FEMALE	CHILD	9	1	YES	NO	NO	NO
1034	8	5	MALE	CHILD	5	0	NO	NO	NO	NO
1035	8	3	FEMALE	CHILD	3	0	YES	NO	NO	NO
1036	8	1	FEMALE	CHILD	1	0	YES	NO	NO	NO
1037	8	60	MALE	PARENT	9	51	NO	NO	NO	NO
1038	9	58	MALE	HH	19	38	YES	YES	YES	YES
1039	9	43	FEMALE	WIFE	19	24	YES	YES	YES	YES
1040	9	13	FEMALE	CHILD	13	0	YES	NO	NO	NO
1041	9	11	MALE	CHILD	11	0	UNKWN	UNKWN	UNKWN	UNKWN
1042	9	7	FEMALE	CHILD	7	0	YES	NO	NO	NO
1043	9	6	FEMALE	CHILD	6	0	YES	NO	NO	NO
1044	9	4	FEMALE	CHILD	4	0	NO	NO	NO	NO
1045	9	46	MALE	OTHER	4	42	YES	NO	NO	NO
1046	9	36	FEMALE	OTHER	4	32	YES	NO	NO	NO
1047	9	12	FEMALE	OTHER	4	8	YES	NO	NO	NO
1048	9	9	MALE	OTHER	4	5	YES	NO	NO	NO
1049	9	7	MALE	OTHER	4	3	YES	NO	NO	NO
1050	9	6	FEMALE	OTHER	4	2	YES	NO	NO	NO
1051	9	4	FEMALE	OTHER	4	0	YES	NO	NO	NO
1052	9	2	FEMALE	OTHER	2	0	YES	NO	NO	NO
1053	9	1	MALE	OTHER	1	1	YES	NO	NO	NO
1054	10	51	MALE	HH	3	48	YES	NO	NO	NO

PAHANG MALAYSIA 1GO

REGNO	HSENO	AGE	SEX	RELHH	YRSIN	YROUT	HELTY	BATRI	WASRI	DNKRI
1055	10	50	FEMALE	WIFE	3	47	YES	NO	NO	NO
1056	10	16	FEMALE	CHILD	3	13	YES	NO	NO	NO
1057	10	11	FEMALE	CHILD	3	8	YES	NO	NO	NO
1058	10	8	MALE	CHILD	3	5	NO	NO	NO	NO
1059	11	42	MALE	HH	5	37	YES	NO	NO	NO
1060	11	31	FEMALE	WIFE	5	26	YES	NO	NO	NO
1061	11	55	MALE	PARENT	5	50	NO	YES	YES	YES
1062	11	10	MALE	CHILD	5	5	YES	NO	NO	NO
1063	11	8	FEMALE	CHILD	5	3	YES	NO	NO	NO
1064	11	17	MALE	FRIEND	4	13	NO	YES	YES	YES
1065	11	19	FEMALE	FRIEND	10	9	YES	YES	YES	YES
1066	11	2	FEMALE	CHILD	2	0	YES	YES	NO	YES
1067	12	40	MALE	HH	UNKWN	UNKWN	NO	NO	NO	NO
1068	12	17	FEMALE	WIFE	2	15	YES	NO	NO	NO
1069	12	6	FEMALE	CHILD	6	0	YES	NO	NO	NO
1070	12	4	FEMALE	CHILD	4	0	YES	NO	NO	NO
1071	12	3	MALE	CHILD	3	0	YES	NO	NO	NO
1072	12	1	MALE	CHILD	1	0	YES	NO	NO	NO
1073	13	21	FEMALE	WIFE	7	14	YES	NO	NO	NO
1074	13	22	FEMALE	OTHER	4	18	YES	NO	NO	NO
1075	13	6	MALE	CHILD	6	0	NO	NO	NO	NO
1076	13	3	FEMALE	CHILD	3	0	YES	NO	NO	NO
1077	13	1	MALE	CHILD	1	0	YES	NO	NO	NO
1078	13	2	MALE	OTHER	2	0	YES	NO	NO	NO
1079	13	28	MALE	CHILD	28	0	YES	NO	NO	NO
1080	13	17	FEMALE	CHILD	17	0	YES	NO	NO	NO
1081	13	11	FEMALE	CHILD	11	0	NO	NO	NO	NO
1082	13	57	MALE	HH	20	37	YES	NO	NO	NO
1083	14	42	MALE	HH	30	12	YES	NO	NO	NO
1084	14	38	FEMALE	WIFE	30	8	YES	NO	NO	NO
1085	14	3	MALE	CHILD	3	0	YES	NO	NO	NO
1086	15	29	MALE	HH	2	27	NO	NO	NO	NO
1087	15	31	FEMALE	WIFE	2	29	YES	NO	NO	NO
1088	15	7	MALE	CHILD	2	5	YES	NO	NO	NO
1089	15	6	FEMALE	CHILD	2	4	YES	NO	NO	NO
1090	15	0	MALE	CHILD	0	0	NO	NO	NO	NO
1091	15	10	FEMALE	OTHER	1	9	YES	NO	NO	NO
1092	16	48	MALE	PARENT	10	38	YES	NO	NO	NO
1093	16	48	FEMALE	WIFE	10	38	YES	NO	NO	NO
1094	16	11	FEMALE	CHILD	11	0	YES	NO	NO	NO
1095	16	7	MALE	CHILD	7	0	UNKWN	UNKWN	UNKWN	UNKWN
1096	16	3	MALE	CHILD	3	0	YES	NO	NO	NO
1097	16	2	MALE	CHILD	2	0	NO	NO	NO	NO
1098	16	42	MALE	OTHER	10	32	YES	NO	NO	NO
1099	16	35	FEMALE	OTHER	10	25	YES	NO	NO	NO
1100	16	46	MALE	OTHER	10	36	YES	NO	NO	NO
1101	16	22	FEMALE	OTHER	10	12	YES	NO	NO	NO
1102	16	41	FEMALE	OTHER	UNKWN	UNKWN	YES	NO	NO	NO
1103	17	41	MALE	HH	7	34	YES	NO	NO	NO
1104	18	30	FEMALE	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1105	18	5	FEMALE	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1106	18	3	FEMALE	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1107	18	56	MALE	HH	30	26	YES	NO	NO	NO
1109	18	40	MALE	UNKWN	30	10	YES	NO	NO	NO

PAHANG MALAYSIA 1GO

REGNO	HSENG	AGE	SEX	RELHH	YRSIN	YROUT	HELTY	BATRI	WASRI	DNKRI
1110	18	25	FEMALE	OTHER	4	21	NO	NO	NO	NO
1111	18	3	MALE	OTHER	3	0	YES	NO	NO	NO
1250	15	55	FEMALE	PARENT	30	25	NO	NO	NO	NO

PAHANG MALAYSIA 1GO

REGNO	FSHRI	SCKBL	BLDDR	KELPN	KELAB	SWNEX	MALIA	OEDEM	SCAR	THKEP
1001	YES	NO	NO	NO	NO	NO	YES	NO	NO	UNKWN
1002	YES	NO	NO	NO	NO	NO	NO	NO	NO	UNKWN
1003	YES	UNKWN	NO	NO	NO	NO	NO	NO	NO	NO
1004	NO	YES	NO	UNKWN	NO	NO	YES	NO	NO	NO
1005	YES	NO	NO	UNKWN	NO	NO	NO	NO	NO	UNKWN
1006	YES	NO	NO	NO	YES	NO	YES	NO	YES	UNKWN
1007	YES	NO	NO	YES	NO	NO	YES	NO	NO	NO
1008	UNKWN	NO	NO	NO	NO	NO	YES	NO	NO	UNKWN
1009	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1010	YES	NO	NO	NO	NO	NO	NO	NO	NO	UNKWN
1011	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO
1012	YES	UNKWN	NO	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1013	NO	UNKWN	NO	UNKWN	UNKWN	NO	YES	NO	NO	UNKWN
1014	NO	NO	NO	UNKWN	UNKWN	NO	YES	NO	NO	UNKWN
1015	NO	UNKWN	NO	UNKWN	NO	NO	YES	NO	NO	YES
1016	YES	NO	NO	NO	NO	NO	YES	NO	NO	UNKWN
1017	YES	NO	NO	NO	NO	NO	YES	NO	NO	NO
1018	NO	NO	NO	UNKWN	UNKWN	NO	YES	NO	NO	NO
1019	NO	NO	NO	UNKWN	UNKWN	NO	YES	NO	NO	NO
1020	NO	NO	NO	UNKWN	UNKWN	NO	YES	NO	NO	UNKWN
1021	NO	UNKWN	NO	UNKWN	UNKWN	NO	YES	NO	NO	UNKWN
1022	NO	UNKWN	NO	UNKWN	UNKWN	NO	YES	NO	NO	NO
1023	NO	NO	NO	NO	NO	NO	YES	NO	NO	UNKWN
1024	NO	NO	NO	YES	NO	YES	YES	NO	YES	UNKWN
1025	NO	YES	NO	YES	YES	NO	YES	NO	YES	NO
1026	NO	NO	YES	NO	NO	NO	YES	NO	NO	UNKWN
1027	NO	NO	NO	NO	NO	NO	YES	NO	NO	UNKWN
1028	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1029	NO	NO	NO	NO	NO	NO	YES	NO	NO	UNKWN
1030	NO	NO	NO	YES	NO	NO	YES	NO	NO	UNKWN
1031	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1032	NO	NO	NO	NO	NO	NO	YES	NO	NO	UNKWN
1033	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1034	NO	NO	YES	NO	NO	NO	YES	NO	NO	UNKWN
1035	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
1036	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO
1037	NO	NO	NO	YES	YES	NO	YES	NO	NO	UNKWN
1038	YES	NO	YES	NO	NO	NO	YES	NO	NO	UNKWN
1039	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO
1040	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
1041	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1042	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
1043	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
1044	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO
1045	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
1046	NO	NO	NO	YES	YES	NO	NO	NO	YES	NO
1047	NO	NO	NO	YES	NO	NO	YES	NO	NO	NO
1048	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1049	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1050	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1051	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1052	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1053	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
1054	NO	NO	NO	NO	NO	NO	YES	NO	YES	UNKWN

PAHANG MALAYSIA 1GO

REGNO	FSHRI	SCKBL	BLDDR	KELPN	KELAB	SWNEX	MALIA	OEDEM	SCAR	THKEP
1055	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
1056	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
1057	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
1058	NO	NO	NO	NO	NO	NO	YES	NO	NO	UNKWN
1059	NO	NO	NO	YES	YES	NO	YES	NO	YES	NO
1060	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1061	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO
1062	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO
1063	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO
1064	YES	YES	NO	NO	NO	NO	YES	NO	NO	NO
1065	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO
1066	NO	NO	NO	UNKWN	UNKWN	NO	YES	NO	NO	NO
1067	NO	NO	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1068	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1069	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1070	NO	UNKWN	UNKWN	NO	NO	NO	NO	NO	NO	NO
1071	NO	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1072	NO	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1073	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
1074	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
1075	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1076	NO	UNKWN	NO	NO	NO	NO	NO	NO	NO	NO
1077	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1078	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
1079	NO	NO	NO	NO	NO	NO	YES	NO	NO	UNKWN
1080	NO	NO	NO	NO	NO	NO	NO	NO	UNKWN	NO
1081	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1082	NO	NO	NO	NO	NO	NO	YES	NO	NO	UNKWN
1083	NO	NO	NO	NO	NO	NO	NO	NO	NO	UNKWN
1084	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
1085	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
1086	NO	YES	YES	YES	NO	NO	YES	NO	NO	NO
1087	NO	NO	YES	NO	NO	NO	YES	NO	NO	NO
1088	NO	NO	YES	NO	NO	NO	NO	NO	NO	UNKWN
1089	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1090	NO	NO	NO	NO	NO	NO	YES	NO	NO	UNKWN
1091	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
1092	NO	NO	NO	YES	YES	YES	NO	NO	NO	NO
1093	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1094	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
1095	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1096	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1097	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1098	NO	NO	NO	NO	NO	NO	YES	NO	NO	UNKWN
1099	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO
1100	NO	NO	NO	NO	NO	YES	YES	YES	NO	NO
1101	NO	NO	NO	NO	NO	NO	NO	UNKWN	UNKWN	UNKWN
1102	NO	NO	NO	NO	NO	NO	NO	UNKWN	UNKWN	UNKWN
1103	NO	NO	NO	NO	NO	NO	YES	NO	NO	UNKWN
1104	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1105	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1106	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1107	NO	NO	YES	YES	YES	NO	YES	NO	NO	UNKWN
1109	NO	NO	NO	NO	NO	NO	NO	NO	YES	NO

PAHANG MALAYSIA 1GO

REGNO	FSHRI	SCKBL	BLDDR	KELPN	KELAB	SWNEX	MALIA	OEDEM	SCAR	THKEP
1110	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
1111	NO	NO	NO	NO	NO	NO	YES	NO	NO	UNKWN
1250	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO

PAHANG MALAYSIA 1GO

REGNO	SPLN	RTLVR	LTLVR	VENAB	INFMTH	MAL	SPMAL	FMMAL	FIL	SPFIL
1001	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1002	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1003	0	YES	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1004	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1005	3	NO	NO	NO	0	NEG	NEG	NEG	NEG	SUPER
1006	0	NO	NO	NO	0	NEG	NEG	NEG	POS	PERIOD
1007	3	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1008	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1009	3	NO	NO	NO	0	NEG	P.F.	RINGS	POS	NEG
1010	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1011	2	NO	NO	NO	0	NEG	NEG	RINGS	UNKWN	NEG
1012	UNKWN	UNKWN	UNKWN	UNKWN	0	NEG	NEG	UNKWN	UNKWN	UNKWN
1013	2	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1014	4	NO	NO	NO	0	POS	P.F.	RINGS	NEG	NEG
1015	0	NO	NO	NO	0	POS	P.V.	RINGS	NEG	NEG
1016	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1017	0	YES	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1018	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1019	2	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1020	3	NO	NO	NO	0	POS	P.F.	GAMETO	NEG	NEG
1021	0	NO	NO	NO	0	POS	P.F.	GAMETO	NEG	NEG
1022	0	YES	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1023	0	YES	YES	NO	0	NEG	NEG	NEG	NEG	NEG
1024	2	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1025	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1026	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1027	2	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1028	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1029	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1030	2	NO	NO	NO	0	NEG	P.F.	GAMETO	NEG	NEG
1031	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1032	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1033	2	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1034	3	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1035	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1036	2	NO	NO	NO	0	POS	P.F.	MIXED	NEG	NEG
1037	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1038	0	YES	NO	YES	0	NEG	NEG	NEG	NEG	NEG
1039	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1040	3	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1041	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1042	0	YES	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1043	2	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1044	0	YES	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1045	0	NO	NO	NO	0	NEG	NEG	NEG	POS	PERIOD
1046	0	NO	NO	NO	0	NEG	NEG	NEG	POS	UNKWN
1047	3	NO	NO	NO	0	POS	P.F.	RINGS	NEG	NEG
1048	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1049	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1050	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1051	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1052	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1053	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1054	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG

PAHANG MALAYSIA 1GO

REGNO	SPLN	RTLVR	LTLVR	VENAB	INFMTH	MAL	SPMAL	FMMAL	FIL	SPFIL
1055	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1056	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1057	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1058	2	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1059	3	NO	NO	UNKWN	0	NEG	NEG	NEG	POS	PERIOD
1060	0	YES	YES	NO	0	NEG	NEG	NEG	NEG	NEG
1061	0	YES	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1062	2	YES	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1063	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1064	0	NO	NO	NO	0	NEG	NEG	NEG	POS	PERIOD
1065	2	YES	NO	NO	0	NEG	NEG	NEG	POS	PERIOD
1066	3	NO	NO	NO	0	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1067	UNKWN	UNKWN	UNKWN	UNKWN	0	NEG	NEG	NEG	NEG	NEG
1068	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1069	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1070	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1071	UNKWN	UNKWN	UNKWN	UNKWN	0	NEG	NEG	NEG	NEG	NEG
1072	UNKWN	UNKWN	UNKWN	UNKWN	0	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1073	1	YES	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1074	0	YES	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1075	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1076	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1077	0	NO	NO	NO	0	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1078	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1079	0	NO	NO	NO	0	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1080	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1081	2	NO	NO	NO	0	POS	P.F.	MIXED	NEG	NEG
1082	2	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1083	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1084	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1085	0	NO	NO	NO	0	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1086	0	NO	NO	NO	0	NEG	NEG	NEG	POS	SUBPER
1087	0	NO	NO	NO	0	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1088	2	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1089	2	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1090	0	NO	NO	NO	9	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1091	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1092	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1093	0	YES	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1094	0	YES	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1095	UNKWN	UNKWN	UNKWN	UNKWN	0	NEG	NEG	NEG	NEG	NEG
1096	1	YES	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1097	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1098	3	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1099	0	YES	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1100	2	NO	NO	NO	0	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1101	UNKWN	UNKWN	UNKWN	UNKWN	0	NEG	NEG	NEG	NEG	NEG
1102	UNKWN	UNKWN	UNKWN	UNKWN	0	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN
1103	3	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1104	UNKWN	UNKWN	UNKWN	UNKWN	0	NEG	NEG	NEG	POS	UNKWN
1105	UNKWN	UNKWN	UNKWN	UNKWN	0	NEG	NEG	NEG	NEG	NEG
1106	UNKWN	UNKWN	UNKWN	UNKWN	0	NEG	NEG	NEG	NEG	NEG
1107	4	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1109	0	NO	NO	NO	0	NEG	NEG	NEG	NEG	NEG

PAHANG MALAYSIA 1GO

REGNO	SPLN	RTLVR	LTLVR	VENAB	INFMTH	MAL	SPMAL	FMMAL	FIL	SPFIL
1110	0	YES	NO	NO	0	POS	P.F.	RINGS	NEG	NEG
1111	2	NC	NO	NO	0	NEG	NEG	NEG	NEG	NEG
1250	2	NO	NO	NO	0	UNKWN	UNKWN	UNKWN	UNKWN	UNKWN

PAHANG MALAYSIA 1GO

REGNO	NOMF	ASC	TRICH	HW
1001	0	POS	POS	POS
1002	0	POS	POS	POS
1003	0	NEG	POS	POS
1004	0	UNKWN	UNKWN	UNKWN
1005	1	POS	POS	POS
1006	5	NEG	POS	POS
1007	0	POS	POS	NEG
1008	0	UNKWN	UNKWN	UNKWN
1009	0	NEG	POS	POS
1010	0	POS	POS	POS
1011	2	POS	POS	POS
1012	UNKWN	UNKWN	UNKWN	UNKWN
1013	0	POS	POS	POS
1014	0	POS	POS	POS
1015	0	POS	NEG	NEG
1016	0	NEG	POS	POS
1017	0	UNKWN	UNKWN	UNKWN
1018	0	NEG	POS	POS
1019	0	NEG	POS	POS
1020	0	POS	POS	POS
1021	0	UNKWN	UNKWN	UNKWN
1022	0	UNKWN	UNKWN	UNKWN
1023	0	UNKWN	UNKWN	UNKWN
1024	0	UNKWN	UNKWN	UNKWN
1025	0	UNKWN	UNKWN	UNKWN
1026	0	UNKWN	UNKWN	UNKWN
1027	0	POS	POS	POS
1028	0	POS	POS	POS
1029	0	NEG	POS	POS
1030	5	POS	POS	POS
1031	0	POS	POS	POS
1032	0	POS	POS	POS
1033	0	POS	POS	NEG
1034	0	POS	POS	POS
1035	0	POS	POS	POS
1036	0	NEG	NEG	NEG
1037	0	POS	POS	POS
1038	0	POS	POS	POS
1039	0	POS	POS	POS
1040	0	POS	POS	POS
1041	UNKWN	UNKWN	UNKWN	UNKWN
1042	0	POS	POS	POS
1043	0	NEG	POS	NEG
1044	0	POS	POS	POS
1045	2	POS	POS	NEG
1046	1	NEG	POS	POS
1047	1	POS	POS	POS
1048	1	POS	POS	POS
1049	1	NEG	POS	POS
1050	0	POS	POS	POS
1051	0	UNKWN	UNKWN	UNKWN
1052	0	NEG	POS	NEG
1053	0	POS	NEG	NEG
1054	0	POS	POS	POS

PAHANG MALAYSIA 1GO

REGNO	NOMF	ASC	TRICH	HW
1055	0	POS	NEG	POS
1056	0	POS	POS	POS
1057	0	POS	POS	POS
1058	0	POS	POS	POS
1059	14	POS	POS	POS
1060	0	POS	POS	POS
1061	0	POS	POS	POS
1062	0	POS	POS	NEG
1063	0	POS	POS	POS
1064	7	UNKWN	UNKWN	UNKWN
1065	7	POS	POS	POS
1066	UNKWN	UNKWN	UNKWN	UNKWN
1067	0	POS	POS	POS
1068	15	POS	POS	POS
1069	0	POS	POS	POS
1070	0	POS	POS	POS
1071	0	UNKWN	UNKWN	UNKWN
1072	UNKWN	UNKWN	UNKWN	UNKWN
1073	UNKWN	POS	POS	POS
1074	UNKWN	POS	POS	POS
1075	UNKWN	POS	POS	POS
1076	UNKWN	UNKWN	UNKWN	UNKWN
1077	UNKWN	UNKWN	UNKWN	UNKWN
1078	0	UNKWN	UNKWN	UNKWN
1079	UNKWN	UNKWN	UNKWN	UNKWN
1080	0	POS	POS	POS
1081	0	POS	POS	NEG
1082	0	POS	POS	POS
1083	0	POS	POS	POS
1084	0	POS	POS	POS
1085	UNKWN	POS	POS	NEG
1086	25	POS	NEG	NEG
1087	UNKWN	POS	POS	NEG
1088	0	POS	POS	POS
1089	0	NEG	POS	POS
1090	UNKWN	UNKWN	UNKWN	UNKWN
1091	0	UNKWN	UNKWN	UNKWN
1092	0	NEG	POS	POS
1093	0	NEG	POS	POS
1094	0	NEG	POS	POS
1095	0	UNKWN	UNKWN	UNKWN
1096	0	NEG	POS	POS
1097	0	POS	POS	NEG
1098	0	POS	POS	POS
1099	0	POS	POS	POS
1100	UNKWN	UNKWN	UNKWN	UNKWN
1101	0	NEG	POS	POS
1102	UNKWN	UNKWN	UNKWN	UNKWN
1103	0	POS	POS	POS
1104	4	UNKWN	UNKWN	UNKWN
1105	0	POS	POS	POS
1106	0	UNKWN	UNKWN	UNKWN
1107	0	POS	POS	POS
1109	0	NEG	POS	POS

PAHANG MALAYSIA 1GO

REGNO	NOMF	ASC	TRICH	HW
1110	0	NEG	POS	POS
1111	0	POS	POS	POS
1250	UNKWN	UNKWN	UNKWN	UNKWN

PAHANG MALAYSIA 1GO

NUMBER	NAME	HIGH	LOW	KNOWN CASES	MEAN	S.D.	COEF.VAR
1	REGNO	1250	1001	111	1057.279	36.745	3.475
2	HSENO	18	1	111	10.063	4.854	48.243
3	AGE	64	0	111	21.225	18.724	88.218
4	SEX	2	1	111	1.495	0.502	33.584
5	RELHH	7	1	107	3.345	1.759	52.592
6	YRSIN	30	0	106	7.877	7.967	101.143
7	YROUT	51	0	106	13.226	15.521	117.352
8	HELTY	2	1	106	1.254	0.437	34.890
9	BATRI	2	1	106	1.726	0.447	25.944
10	WASRI	2	1	106	1.792	0.407	22.732
11	DNKRI	2	1	106	1.726	0.447	25.944
12	FSHRI	2	1	105	1.847	0.361	19.544
13	SCKBL	2	1	96	1.958	0.200	10.257
14	BLDDR	2	1	102	1.892	0.31	16.474
15	KELPN	2	1	91	1.879	0.32	17.443
16	KELAB	2	1	94	1.925	0.263	13.707
17	SWNEX	2	1	102	1.970	0.169	8.616
18	SCRSW	2	2	101	2.000	0.000	0.000
19	SCRPN	2	2	101	2.000	0.000	0.000
20	CHUIA	2	2	99	2.000	0.000	0.000
21	MALIA	2	1	102	1.392	0.490	35.243
22	ELEPH	2	2	99	2.000	0.000	0.000
23	OEDEM	2	1	100	1.989	0.100	5.025
24	SCAR	2	1	99	1.929	0.257	13.354
25	HYCLE	2	2	96	2.000	0.000	0.000
26	THKEP	2	1	68	1.985	0.121	6.108
27	SCBRE	2	2	79	2.000	0.000	0.000
28	SPLN	4	0	100	0.809	1.203	148.578
29	RTLVR	2	1	100	1.819	0.386	21.215
30	LTLVR	2	1	100	1.979	0.140	7.106
31	VENAB	2	1	99	1.989	0.100	5.050
32	YEAR	80	80	111	80.000	0.000	0.000
33	MONTH	5	5	111	5.000	0.000	0.000
34	VILNO	1	1	111	1.000	0.000	0.000
35	INFMTH	9	0	110	0.081	0.858	1048.808
36	MAL	1	0	100	0.079	0.272	340.824
37	SPMAL	2	0	100	0.109	0.345	313.723
38	FMMAL	3	0	99	0.181	0.577	317.837
39	FIL	1	0	98	0.091	0.290	316.082
40	SPFIL	2	0	97	0.123	0.462	373.676
41	NOMF	25	0	95	0.957	3.476	362.890
42	SCHIS	0	0	83	0.000	0.000	0.000
43	ASC	1	0	83	0.759	0.430	56.686
44	TRICH	1	0	83	0.939	0.239	25.472
45	HW	1	0	83	0.831	0.376	45.318
46	COPT	2	0	84	0.119	0.450	378.022
47	FT	0	0	111	0.000	0.000	0.000

The Summary Statistics just presented may be useful for error checking, but they do not by themselves permit much statistical analysis. In order to be able to perform various statistical tests it is essential to be able to obtain such Summary Statistics for various selected groups. For example, we may wish to compare young with old, males with females, sick with well, and so on. CCSS and many other systems permit selection of groups within the data by means of logical criteria as, for example, SEX = 1 (male). We may also specify logical combinations such as "Males under 40" or "Males who have had malaria", and so on.

To illustrate the listing of Summary Statistics for subgroups, the next 2 pages give statistics for all variables for the "LOCAL" residents, as contrasted with the "MIGRANT" group. For a first experiment, LOCAL was defined to be those with years residing elsewhere (YROUT) equal to 0, i.e. YROUT =0. This even includes the baby as a LOCAL. MIGRANT is then defined to be all those who have spent at least 1 year outside the village, i.e. YROUT 0. A comparison of the 2 sets of statistics may show some interesting differences. For example, one immediately obtains the percentage 37.8% unhealthy for the locals as opposed to 18.8% "unhealthy" among those who have been outside the village 1 year or more. A test could, of course, be performed if desired, to determine statistical significance.

As another illustration, Summary Statistics for those who have had malaria (MALIA = 1) are given. In these data, we see that the Hackett Spleen value ranges from 0 to 4, with an overall mean of 0.81. For the MALIA+ cases, however, the mean of SPLN = 0.97, contrasting with the mean for the MALIA- cases (not shown) of 0.55. Again, a statistical test may be used to determine significance of this difference.

Such statistics for different groups are frequently used for a t-test between two groups or for analysis of variance among many groups. CCSS does not automatically perform these tests, but they can be done on a calculator once the basic means and S.D.s have been found by CCSS. Some other systems may perform these tests automatically. However, there are good reasons for not providing too many tests automatically. The chief problem is that the investigator may too easily accept a value as significant when the test is not even appropriate. This is a good argument in favor of not automatically providing such information. One should consult a statistician before jumping to conclusions. One of the greatest hazards is that of looking at test after test until one is finally found which is theoretically significant. However, if the non-significant tests are taken into account, then the result will usually turn out to be non-significant overall. (Ignoring this problem results in one of the more common fallacies in statistics, discussed by I. J. Good, International Encyclopedia of Statistics, 1978, page 341, "Statistical Fallacies".)

LOCAL 7E0

NUMBER	NAME	HIGH	LOW	KNOWN CASES	MEAN	S.D.	COEF.VAR
1	REGNC	1111	1004	39	1055.820	28.855	2.732
2	HSENO	16	1	39	10.025	4.307	42.961
3	AGE	29	0	39	5.820	6.500	111.682
4	SEX	2	1	39	1.487	0.506	34.049
5	RELHH	6	1	39	3.410	1.163	34.115
6	YRSIN	29	0	39	5.820	6.500	111.682
7	YROUT	0	0	39	0.000	0.000	0.000
8	HELTY	2	1	37	1.378	0.491	35.670
9	BATRI	2	1	37	1.783	0.417	23.396
10	WASRI	2	1	37	1.972	0.164	8.332
11	DNKRI	2	1	37	1.783	0.417	23.396
12	FSHRI	2	2	37	2.000	0.000	0.000
13	SCKBL	2	1	30	1.966	0.182	9.283
14	BLDDR	2	1	34	1.882	0.327	17.373
15	KELPN	2	2	27	2.000	0.000	0.000
16	KELAB	2	2	29	2.000	0.000	0.000
17	SHNEX	2	2	35	2.000	0.000	0.000
18	SCRSW	2	2	34	2.000	0.000	0.000
19	SCRPN	2	2	34	2.000	0.000	0.000
20	CHUIA	2	2	33	2.000	0.000	0.000
21	MALIA	2	1	35	1.342	0.481	35.863
22	ELEPH	2	2	34	2.000	0.000	0.000
23	OEDEM	2	2	35	2.000	0.000	0.000
24	SCAR	2	2	34	2.000	0.000	0.000
25	HYCLE	2	2	35	2.000	0.000	0.000
26	THKEP	2	1	25	1.959	0.200	10.204
27	SCBRE	2	2	31	2.000	0.000	0.000
28	SPLN	4	0	35	0.828	1.248	150.643
29	RTLVR	2	1	35	1.857	0.355	19.117
30	LTLVR	2	2	35	2.000	0.000	0.000
31	VENAB	2	2	35	2.000	0.000	0.000
32	YEAR	80	80	39	80.000	0.000	0.000
33	MONTH	5	5	39	5.000	0.000	0.000
34	VILNO	1	1	39	1.000	0.000	0.000
35	INFMTH	9	0	38	0.236	1.459	616.441
36	MAL	1	0	32	0.187	0.396	211.497
37	SPMAL	2	0	32	0.218	0.490	224.385
38	FMMAL	3	0	32	0.375	0.870	232.178
39	FIL	0	0	32	0.000	0.000	0.000
40	SPFIL	0	0	32	0.000	0.000	0.000
41	NOMF	0	0	30	0.000	0.000	0.000
42	SCHIS	0	0	24	0.000	0.000	0.000
43	ASC	1	0	24	0.708	0.464	65.549
44	TRICH	1	0	24	0.916	0.282	30.799
45	HW	1	0	24	0.708	0.464	65.549
46	COPT	2	0	25	0.079	0.399	499.999
47	RT	0	0	39	0.000	0.000	0.000

MIGRANT 760

NUMBER	NAME	HIGH	LOW	KNOWN CASES	MEAN	S.D.	COEF.VAR
1	REGNO	1250	1001	72	1058.069	40.567	3.834
2	HSENO	18	1	72	10.083	5.156	51.134
3	AGE	64	1	72	29.569	17.878	60.461
4	SEX	2	1	72	1.500	0.503	33.567
5	RELHH	7	1	68	3.308	2.031	61.397
6	YRSIN	30	1	67	9.074	8.526	93.963
7	YROUT	51	1	67	20.925	14.823	70.839
8	HELTY	2	1	69	1.188	0.393	33.145
9	BATRI	2	1	69	1.695	0.463	27.334
10	WASRI	2	1	69	1.695	0.463	27.334
11	DNKRI	2	1	69	1.695	0.463	27.334
12	FSHRI	2	1	68	1.764	0.427	24.215
13	SCKBL	2	1	66	1.954	0.209	10.738
14	BLDDR	2	1	68	1.897	0.306	16.137
15	KELPN	2	1	64	1.828	0.380	20.800
16	KELAB	2	1	65	1.892	0.312	16.509
17	SWNEX	2	1	67	1.955	0.208	10.657
18	SCRSW	2	2	67	2.000	0.000	0.000
19	SCRPN	2	2	67	2.000	0.000	0.000
20	CHUIA	2	2	66	2.000	0.000	0.000
21	MALIA	2	1	67	1.417	0.496	35.047
22	ELEPH	2	2	65	2.000	0.000	0.000
23	OEDEM	2	1	65	1.984	0.124	6.249
24	SCAR	2	1	65	1.892	0.312	16.509
25	HYCLE	2	2	61	2.000	0.000	0.000
26	THKEP	2	2	43	2.000	0.000	0.000
27	SCBRE	2	2	48	2.000	0.000	0.000
28	SPLN	4	0	65	0.799	1.188	148.560
29	RTLVR	2	1	65	1.799	0.403	22.395
30	LTLVR	2	1	65	1.969	0.174	8.837
31	VENAB	2	1	64	1.984	0.125	6.299
32	YEAR	80	80	72	80.000	0.000	0.000
33	MONTH	5	5	72	5.000	0.000	0.000
34	VILNO	1	1	72	1.000	0.000	0.000
35	INFMTH	0	0	72	0.000	0.000	0.000
36	MAL	1	0	68	0.029	0.170	578.727
37	SPMAL	1	0	68	0.058	0.237	402.973
38	FMMAL	2	0	67	0.089	0.336	375.490
39	FIL	1	0	66	0.136	0.345	253.589
40	SPFIL	2	0	65	0.184	0.555	301.165
41	NOMF	25	0	65	1.399	4.137	295.534
42	SCHIS	0	0	59	0.000	0.000	0.000
43	ASC	1	0	59	0.779	0.418	53.617
44	TRICH	1	0	59	0.949	0.221	23.344
45	HW	1	0	59	0.881	0.326	37.004
46	CORT	2	0	59	0.135	0.471	348.016
47	RT	0	0	72	0.000	0.000	0.000

MALIA + 21E1

NUMBER	NAME	HIGH	LOW	KNOWN CASES	MEAN	S.D.	COEF.VAR
1	REGNO	1111	1001	62	1049.548	32.102	3.058
2	HSENO	18	1	62	9.096	4.871	53.548
3	AGE	64	0	62	21.580	19.921	92.312
4	SEX	2	1	62	1.403	0.494	35.243
5	RELHH	7	1	62	3.112	1.802	57.907
6	YRSIN	30	0	62	7.693	7.841	101.927
7	YROUT	51	0	62	13.870	15.780	113.768
8	HELTY	2	1	62	1.354	0.482	35.603
9	BATRI	2	1	62	1.677	0.471	28.095
10	WASRI	2	1	62	1.790	0.410	22.923
11	DNKRI	2	1	62	1.677	0.471	28.095
12	FSHRI	2	1	61	1.885	0.321	17.046
13	SCKBL	2	1	58	1.931	0.255	13.236
14	BLDDR	2	1	62	1.903	0.298	15.661
15	KELPN	2	1	52	1.826	0.382	20.909
16	KELAB	2	1	54	1.907	0.292	15.339
17	SWNEX	2	1	62	1.967	0.178	9.052
18	SCRSW	2	2	61	2.000	0.000	0.000
19	SCRPN	2	2	61	2.000	0.000	0.000
20	CHUIA	2	2	60	2.000	0.000	0.000
21	MALIA	1	1	62	1.000	0.000	0.000
22	ELEPH	2	2	62	2.000	0.000	0.000
23	OEDEM	2	1	62	1.983	0.127	6.401
24	SCAR	2	1	62	1.919	0.274	14.302
25	HYCLE	2	2	59	2.000	0.000	0.000
26	THKEP	2	1	35	1.971	0.169	8.574
27	SCBRE	2	2	45	2.000	0.000	0.000
28	SPLN	4	0	62	0.967	1.305	134.883
29	RTLVR	2	1	62	1.870	0.337	18.064
30	LTLVR	2	1	62	1.967	0.178	9.052
31	VENAB	2	1	61	1.983	0.128	6.454
32	YEAR	80	80	62	80.000	0.000	0.000
33	MONTH	5	5	62	5.000	0.000	0.000
34	VILNO	1	1	62	1.000	0.000	0.000
35	INFMTH	9	0	62	0.145	1.143	787.400
36	MAL	1	0	56	0.107	0.312	291.287
37	SPMAL	2	0	56	0.160	0.416	259.326
38	FMMAL	3	0	56	0.232	0.632	272.331
39	FIL	1	0	56	0.089	0.287	322.264
40	SPFIL	2	0	56	0.125	0.469	375.620
41	HOMF	25	0	55	1.345	4.406	327.495
42	SCHIS	0	0	46	0.000	0.000	0.000
43	ASC	1	0	46	0.760	0.431	56.680
44	TRICH	1	0	46	0.956	0.206	21.555
45	HW	1	0	46	0.826	0.383	46.390
46	CORT	2	0	46	0.152	0.514	338.423
47	RT	0	0	62	0.000	0.000	0.000

B. Tables

In the analysis of survey data, we are often more interested in counts of cases falling into certain categories than in continuous measurements. For example, we may wish to find out if older persons show more evidence, or more often show evidence, of disease than younger persons. We therefore count the numbers of sick and well in both the young and the old age groups. Such counting is done automatically by CCSS and other computer systems in the form of tables.

The simplest form of table is the histogram, which is just a one-dimensional table, or a single column of cells, with counts in each cell. Since these are so simple, they are usually produced with a pictorial representation of bars, to make the presentation more graphic. With a histogram one must specify the intervals, as, for example, age intervals, into which the data are to be classified and counted, as age 0-9, 10-19, 20-29, and so on. One must also specify the subgroup for which the data are to be counted, if desired. For example, a histogram of age for males only could be produced.

CCSS also has a BARGRAPH option which records all of the codes for a variable which occur in the desired subgroup and counts how many times each code occurs.

By all odds the most useful table is the two-way table, with rows and columns, and counts in each cell. Here, one must specify (1) the row intervals and variable, (2) the column intervals and variable, and (3) the subgroup desired. On the following pages are first, a 2-way table showing the age/sex breakdown of the study population, and second, a table showing the Malaria positives by age.

A convenient addition to the 2-way table is a set of percent tables, showing the percentages for each table by (1) row, (2) column, and (3) over all cells.

For 3-way and higher dimensional tables, CCSS provides what are called Nested Tables, which is one big table displaying all combinations of the specified variables and intervals. NAMRU-2 uses, instead, a true 3-way tabling program in its own package called PORTSTAT. Tables up to 8 x 8 x 8 are produced with this program.

PAHANG MALAYSIA 1GO

		SEX MALE	FEMALE	TOTAL	UNKWN
	0-	25	21	46	0
	10-	4	12	16	0
	20-	5	8	13	0
AGE	30-	1	6	7	0
	40-	12	5	17	0
	50-	9	3	12	0
	TOTAL	56	55	111	*****
	UNKWN	0	0	*****	0

CASES IN TABLE= 111 KNOWNS + 0 UNKNOWNS

CHI SQUARE = 14.48607638 DEG. FREEDOM = 5.

ROW PERCENT TABLE FOR...TABLE 1

PAHANG MALAYSIA 1GO

		SEX		
		MALE	FEMALE	TOTAL
AGE	0-	54.34	45.65	100.00
	10-	25.00	75.00	100.00
	20-	38.46	61.53	100.00
	30-	14.28	85.71	100.00
	40-	70.58	29.41	100.00
	50-	75.00	25.00	100.00
	TOTAL	50.45	49.54	100.00

COL PERCENT TABLE FOR...TABLE 1

PAHANG MALAYSIA 1GO

		SEX		
		MALE	FEMALE	TOTAL
AGE	0-	44.64	38.18	41.44
	10-	7.14	21.81	14.41
	20-	8.92	14.54	11.71
	30-	1.78	10.90	6.30
	40-	21.42	9.09	15.31
	50-	16.07	5.45	10.81
	TOTAL	100.00	100.00	100.00

TOTAL PERCENT TABLE FOR...TABLE 1

PAHANG MALAYSIA 1GO

		SEX		
		MALE	FEMALE	TOTAL
AGE	0-	22.52	18.91	41.44
	10-	3.60	10.81	14.41
	20-	4.50	7.20	11.71
	30-	0.90	5.40	6.30
	40-	10.81	4.50	15.31
	50-	8.10	2.70	10.81
	TOTAL	50.45	49.54	100.00

PAHANG MALAYSIA 1G0

		NEG	MAL POS	TOTAL	UNKWN
	0-	36	5	41	5
	10-	13	2	15	1
	20-	11	1	12	1
AGE	30-	6	0	6	1
	40-	15	0	15	2
	50-	11	0	11	1
	TOTAL	92	8	100	*****
	UNKWN	0	0	*****	0

CASES IN TABLE= 100 KNOWNS + 11 UNKNOWNS

CHI SQUARE = 4.34451104 DEG. FREEDOM = 5.

ROW PERCENT TABLE FOR...TABLE 2

PAHANG MALAYSIA 1GO

		MAL		
		NEG	POS	TOTAL
	0-	87.80	12.19	100.00
	10-	86.66	13.33	100.00
AGE	20-	91.66	8.33	100.00
	30-	100.00	0.00	100.00
	40-	100.00	0.00	100.00
	50-	100.00	0.00	100.00
	TOTAL	92.00	8.00	100.00

COL PERCENT TABLE FOR...TABLE 2

PAHANG MALAYSIA 1GO

		MAL		
		NEG	POS	TOTAL
	0-	39.13	62.50	41.00
	10-	14.13	25.00	15.00
AGE	20-	11.95	12.50	12.00
	30-	6.52	0.00	6.00
	40-	16.30	0.00	15.00
	50-	11.95	0.00	11.00
	TOTAL	100.00	100.00	100.00

TOTAL PERCENT TABLE FOR...TABLE 2

PAHANG MALAYSIA 1GO

		MAL		
		NEG	POS	TOTAL
	0-	36.00	5.00	41.00
	10-	13.00	2.00	15.00
AGE	20-	11.00	1.00	12.00
	30-	6.00	0.00	6.00
	40-	15.00	0.00	15.00
	50-	11.00	0.00	11.00
	TOTAL	92.00	8.00	100.00

In the statistical analysis of tables of counts, or contingency tables, the chi-squared test is often used. The CCSS system, therefore, gives the chi-squared value for each table, if desired. Some computer systems give other statistics. However, such statistics should not be used without a careful analysis of the problem to see if the test is actually appropriate. For paired data, for example, in a 2 x 2 table the chi-squared test is not appropriate, and McNemar's test should be used (9).

For the chi-squared test with a 2 x 2 table, the Yates Correction is often used, but examination of hundreds of cases in the range where this correction would be needed indicates that this correction is almost always over-conservative. For practical work, the following alternative to the Yates Correction is preferable. Use the chi-squared without the correction if the total number of cases is 40 or more. If there are fewer cases, or if the result is borderline, use Fisher's Exact Test. Fisher's test has the advantage of being usable either as a one-sided test, in case only one direction of variation is possible, or as a two-sided test, in which case the simplest approach is to double the one-sided value.

In the statistical analysis of survey data, we are frequently faced with possible simultaneous interaction among 3 or more variables. If the analysis is done for all possible pairs of variables involved, say 3 pairs for a total of 3 variables, interactions among all three variables may be masked, or confounded. Examples occur in which no pairwise chi-squared value is significant, yet the 3-way interaction is highly significant. A practical method for testing such interactions for significance is that of Lancaster (7). It has the disadvantage of being biased if the data are highly skewed. For example, in a 2 x 2 x 2 table of sex, age, and health counts, if there are disproportionally many old females and young males in the sample, so that the AGE x SEX 2 x 2 table is highly significant, the method may give a biased value for the 3-way interaction. Nevertheless, the Lancaster method is a highly useful and practical tool and may be applied to many dimensions.

Where the skewing of the data may cast doubt on the multi-dimensional chi-squared analysis, the log-linear method may be used to test for significance of higher interactions (8). Programs are available at NAMRU-2 and elsewhere to carry out this type of analysis.

Regression methods are sometimes applied to count data, possibly because powerful programs for multiple regression are widely available. It is possible to apply them to tables of count data, but this is less natural than the direct approach to discrete data described above.

C. Scattergrams

In survey data, continuous variables, like age, are still often conveniently treated as discrete, by breaking the age into groups, such as YOUNG (0 - 19) and OLD (20+). The reason for this is two-fold. First, the other variables under study may be discrete, as SICK vs. WELL. Second, and more fundamentally, we must be careful to consider just what question we are asking, before we carry out a statistical test and get an answer. Otherwise, we may get the answer to the wrong question. If we really want to know what the effect is of, say, 1 more year of age, as a matter of intensity, then we should, of course, treat the variable as continuous and use methods, such as regression, to study it. However, such data as age are often either highly inaccurate or of little importance compared to the category such as YOUNG vs. OLD. Therefore, the discrete variable approach should be seriously considered in the analysis of survey data before turning to continuous methods.

When we are interested in the analysis of continuous data such as fever temperature, or worm counts (usually taken as the Logarithm), the first step should be to prepare a scattergram showing the interaction between the two variables. CCSS does this automatically and also prints the linear regression equation and the correlation coefficient. Scattergrams should also be prepared for various subgroups if the effects seem to vary with subgroups. Where we are really concerned with quantitative change, as opposed to qualitative, discrete, change, then multiple regression methods may be appropriate.

On the following page there is a scattergram produced by CCSS which shows the interaction between the variables AGE and SPLN (Spleen), restricted to the group with a positive spleen. The correlation coefficient R = 0.11 is not significant and it appears that there is no particularly strong relationship between age and spleen size for this group.

C. Scattergrams

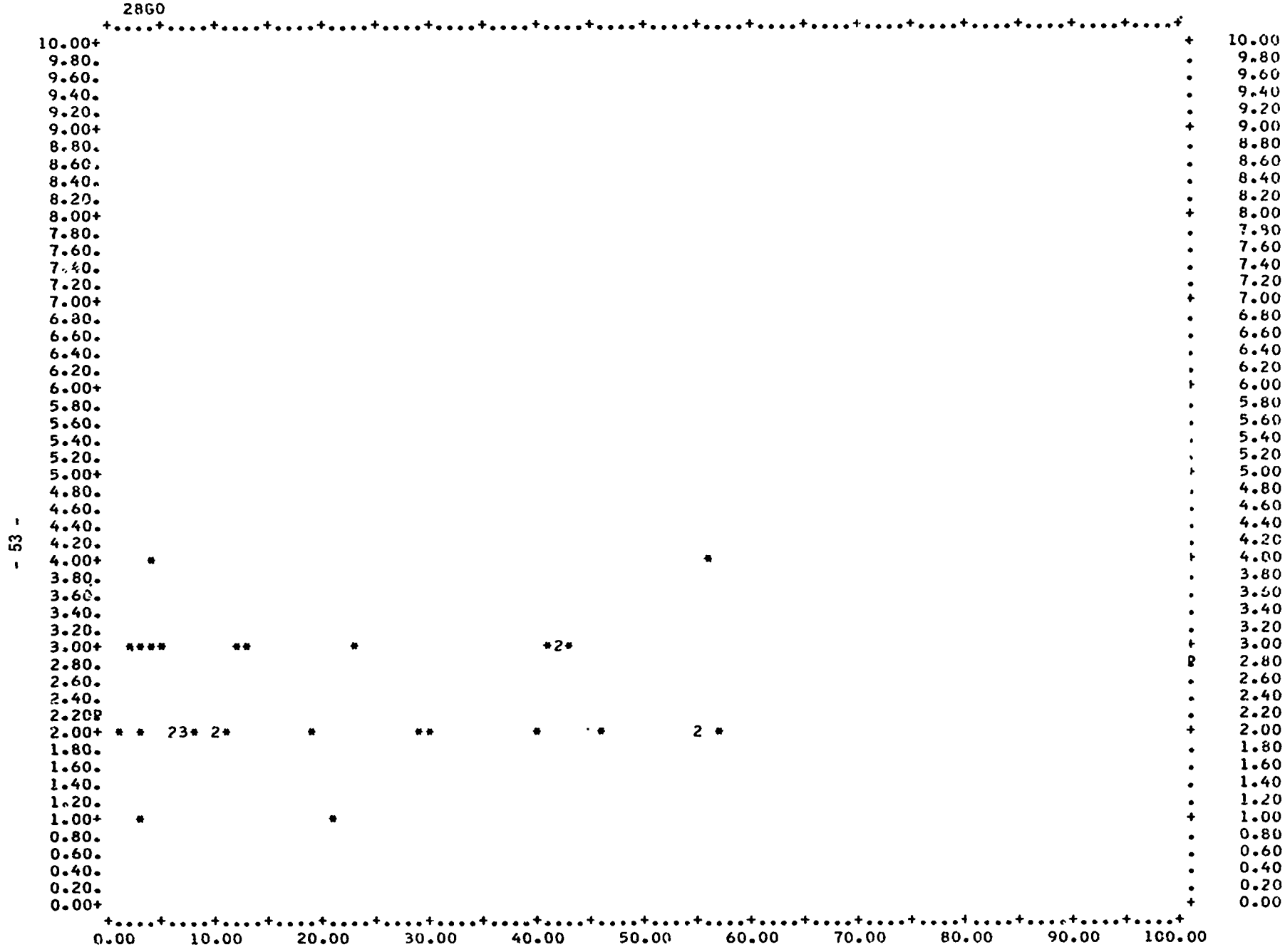
In survey data, continuous variables, like age, are still often conveniently treated as discrete, by breaking the age into groups, such as YOUNG (0 - 19) and OLD (20+). The reason for this is two-fold. First, the other variables under study may be discrete, as SICK vs. WELL. Second, and more fundamentally, we must be careful to consider just what question we are asking, before we carry out a statistical test and get an answer. Otherwise, we may get the answer to the wrong question. If we really want to know what the effect is of, say, 1 more year of age, as a matter of intensity, then we should, of course, treat the variable as continuous and use methods, such as regression, to study it. However, such data as age are often either highly inaccurate or of little importance compared to the category such as YOUNG vs. OLD. Therefore, the discrete variable approach should be seriously considered in the analysis of survey data before turning to continuous methods.

When we are interested in the analysis of continuous data such as fever temperature, or worm counts (usually taken as the Logarithm), the first step should be to prepare a scattergram showing the interaction between the two variables. CCSS does this automatically and also prints the linear regression equation and the correlation coefficient. Scattergrams should also be prepared for various subgroups if the effects seem to vary with subgroups. Where we are really concerned with quantitative change, as opposed to qualitative, discrete, change, then multiple regression methods may be appropriate.

On the following page there is a scattergram produced by CCSS which shows the interaction between the variables AGE and SPLN (Spleen), restricted to the group with a positive spleen. The correlation coefficient R = 0.11 is not significant and it appears that there is no particularly strong relationship between age and spleen size for this group.

SCATTERGRAM

PLOT 1.SPLEEN + 34 POINTS HORIZ. AXIS= AGE VERT. AXIS= SPLN



Y= 2.3823+ 0.0041(X- 21.2058)
R= 0.1147 XSD= 19.0673 YSD= 0.6969 SYX= 0.7030

X. Integrating Statistical Analysis and Data Processing within the Research Process

The statistician, computer personnel, and investigators should form a team to analyze the data if the research program involving survey data is to succeed. If there is a weakness at any point the entire project may be jeopardized. If the points covered in sections I, II, and III of this Guide are followed, the conditions are then favorable for a successful project. In particular, the statistician and the computer personnel must be brought into the project at the beginning, because it is often impossible to salvage a project which has started from the wrong foundation.

XI. Pitfalls to be Avoided

As pointed out in Section I of this Guide, carrying out a research project is "like walking a mountain path", for it is easy to lose one's way at any point. Once the study design has been fixed, procedures must be set up to catch errors at every point, as has been stressed in this Guide. A pilot study is very valuable when possible, to provide insight into the reasonableness of the assumptions and the data gathering procedures. In any case the statistician should check the data at various stages to ensure that nothing goes wrong, so as to avoid the consequences of Murphy's Law, stated earlier: "If anything can go wrong, it will!"

REFERENCES

1. Fox, John P., Hall, Carrie E. and Elveback, Lila R. (1970) Epidemiology, Man and Disease, New York, Macmillan.

2. Lilienfeld, Abraham M. and Lilienfeld, David E. (1980) Foundations of Epidemiology, Second Ed., New York, Oxford University Press.

3. Leech, F.B., and Sellers, K.C. (1979) Statistical Epidemiology in Veterinary Science, New York, Macmillan.

4. Dennis, D.T. (1980) Report on an Epidemiologic Field Study at Kampung Kuala Koyan, Pahang 19-21 May 1980, Kuala Lumpur, WHO.

5. Kronmal, R. (1974) A Conversational Computer Statistical System, Seattle, University of Washington.

6. Good, I.J. (1978) Fallacies, Statistical, in International Encyclopedia of Statistics, New York, Free Press.

7. Lancaster, H.O. (1951) Complex Contingency Tables Treated by Partition of X, J. Royal Statist. Soc. Ser. B, No. 13, pp. 242-249

8. Bishop, Yvonne, M.M., Fienberg, Stephen E., and Holland, Paul W. (1975) Discrete Multivariate Analysis: Theory and Practice, Cambridge, Mass., MIT Press.

9. Armitage, P. (1971) Statistical Methods in Medical Research, New York, Wiley.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: SP-NAMRU-2-46 AD-A123575 / 2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): A Guide to Handling Biomedical Data	5. TYPE OF REPORT & PERIOD COVERED: Special Publication
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Richard See	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: U.S. Naval Medical Research Unit No. 2, APO San Francisco, CA 96528	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: M0106PN.01.1002
11. CONTROLLING OFFICE NAME AND ADDRESS: Commanding Officer, Naval Medical Research and Development Command, National Naval Medical Center, Bethesda, MD 20814	12. REPORT DATE: 1981 13. NUMBER OF PAGES: 55
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): --	15. SECURITY CLASS. (of this report): Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this document is unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

A Special Publication of the U.S. Naval Medical Research Unit No. 2

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Biomedical Data
Handling Biomedical Data
Guide

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This special publication is a guide to handling biomedical data collected during surveys and in carrying out certain biomedical investigations. The system has been used extensively by NAMRU-2 and has been found to be highly successful. Information is presented on the various stages in data processing inlcuding, codesheets, data recording transciption, data entry, presentation and checking and statistical analysis. Certain pitfalls to be avoided are also presented.

DD FORM 1 JAN 73 1473

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ADMINISTRATIVE INFORMATION

This work was supported through funds provided by the U.S. Naval Medical Research and Development Command, Navy Department for Work Unit No. M0106PN.01.1002.

W.H. SCHROEDER
CAPT MSC USN